State of the art of the Automated Similarity Judgment Program

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

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

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

(See further Brown et al. 2008)

THE MEASUREMENT OF STABILITIES

- Count proportions of matches for pairs of words with similar meanings among languages within genera
- Add corrections for chance agreement
- Use weighted means

(See further Holman et al. 2008)

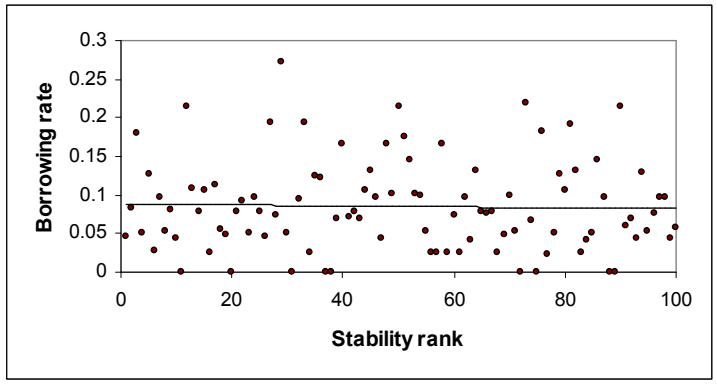
RESULTS FOR STABILITIES OF SWADESH LIST ITEMS

(Starred members are the ones that were selected for the shorter 40-item list)

Rank	# In list	Meaning	Stability	Rank	# In list	Meaning	Stability
1	22	*louse	42.8	51	89	yellow	22.5
2	12	*two	39.8	52	20	bird	21.8
3	75	*water	37.4	53	38	head	21.7
4	39	*ear	37.2	54	79	earth	21.7
5	61	*die	36.3	55	46	foot	21.6
6	1	*I	35.9	56	91	black	21.6
7	53	*liver	35.7	57	42	mouth	21.5
8	40	*eye	35.4	58	88	green	21.1
9	48	*hand	34.9	59	60	sleep	21.0
10	58	*hear	33.8	60	7	what	20.7
11	23	*tree	33.6	61	26	root	20.5

12	19	*fish	33.4	62	45	claw	20.5
13	100	*name	32.4	63	56	bite	20.5
14	77	*stone	32.1	64	83	ash	20.3
15	43	*tooth	30.7	65	87	red	20.2
16	51	*breasts	30.7	66	55	eat	20.0
17	2	*you	30.6	67	33	egg	19.8
18	85	*path	30.2	68	6	who	19.0
19	31	*bone	30.1	69	99	dry	18.9
20	44	*tongue	30.1	70	37	hair	18.6
21	28	*skin	29.6	71	81	smoke	18.5
22	92	*night	29.6	72	8	not	18.3
23	25	*leaf	29.4	73	4	this	18.2
24	76	rain	29.3	74	24	seed	18.2
25	62	kill	29.2	75	16	woman	17.9
26	30	*blood	29.0	76	98	round	17.9
27	34	*horn	28.8	77	14	long	17.4
28	18	*person	28.7	78	69	stand	17.1
29	47	*knee	28.0	79	97	good	16.9
30	11	*one	27.4	80	17	man	16.7
31	41	*nose	27.3	81	94	cold	16.6
32	95	*full	26.9	82	29	flesh	16.4
33	66	*come	26.8	83	50	neck	16.0
34	74	*star	26.6	84	71	say	16.0
35	86	*mountain	26.2	85	84	burn	15.5
36	82	*fire	25.7	86	35	tail	14.9
37	3	*we	25.4	87	78	sand	14.9
38	54	*drink	25.0	88	5	that	14.7
39	57	*see	24.7	89	65	walk	14.4
40	27	bark	24.5	90	68	sit	14.3
41	96	*new	24.3	91	10	many	14.2
42	21	*dog	24.2	92	9	all	14.1
43	72	*sun	24.2	93	59	know	14.1
44	64	fly	24.1	94	80	cloud	13.9
45	32	grease	23.4	95	63	swim	13.6
46	73	moon	23.4	96	49	belly	13.5
47	70	give	23.3	97	13	big	13.4
48	52	heart	23.2	98	93	hot	11.6
49	36	feather	23.1	99	67	lie	11.2
50	90	white	22.7	100	15	small	6.3

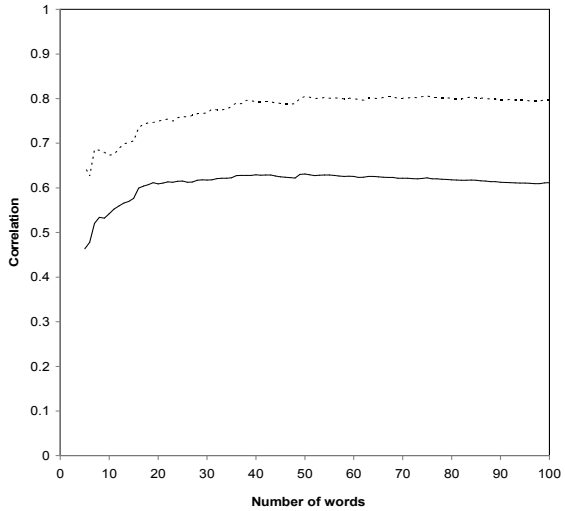
CORRELATING STABILITY AND BORROWABILITY



Potential explanations for the absence of a correlation:

- 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).

CORRELATION BETWEEN DISTANCES IN THE AUTOMATED APPROACH AND OTHER CLASSIFICATIONS AS A FUNCTION OF LIST LENGTHS



Top curve: Ethnologue (correlation method: Goodman-Kruskal gamma)
Bottom curve: WALS (Pearson product-moment correlation)

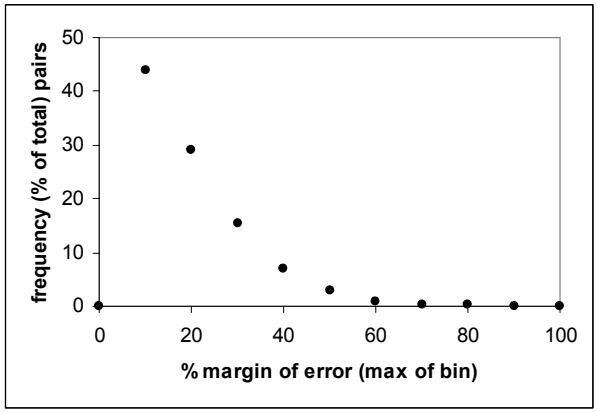
WEIGHTED LEVENSHTEIN DISTANCES

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)

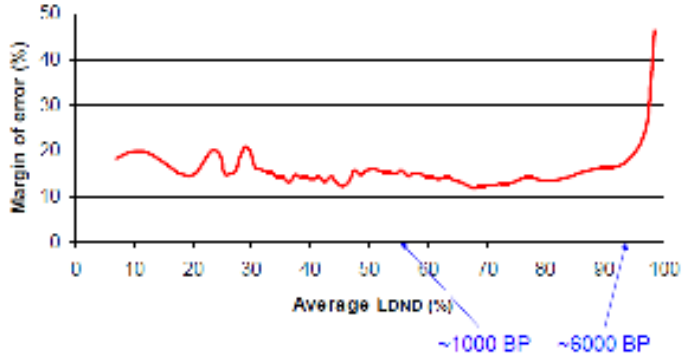
PERFORMANCE OF CLASSIFICATION

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		

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 hat are within the same 1% interval.
- y-axis: the margin of error estimated as the average of the differences between he (logarithms of) the two largest distances for the set of triplets in the interval ivided by the (logarithm) of the average of these two largest distances.

THE REVISED GLOTTOCHRONOLOGICAL FORMULA

Standard formula:

$$\log(SIM) = [2\log(R)]T$$

New formula taking into account inherent variability within languages

$$\log(SIM) = [2\log(R)]\ T + \log(SIM')$$

SIM = observed similarity = 1-LDND; SIM' = baseline similarity at time 0; R = retention rate; T = time in millenia
Preliminary results of calibration: R = .81 (slope of the line) ; SIM' = .68 (the intercept).
Final formula:

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

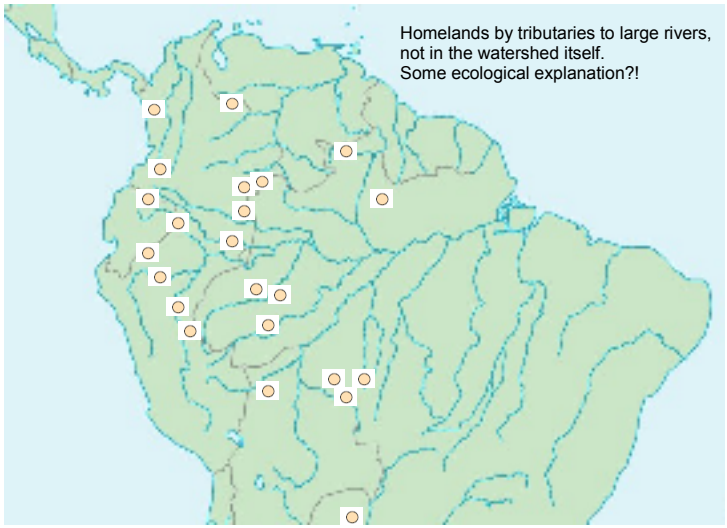
SOME EXAMPLES OF RESULTS

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

A QUANTITATIVE IMPLEMENTATION OF THE HOMELAND IDENTIFICATION PROCEDURE

- 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.

AN EXAMPLE: HOMELANDS OF SOUTH AMERICAN FAMILIES



SELECTED REFERENCES

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the World's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals* 61.4: 285-308.

Diamond, Jared and Peter Bellwood. 2003. Farmers and their languages: the first expansions. *Science* 300: 597-603.

Harlan, Jack R. 1971. Agricultural Origins: Centers and Noncenters. *Science* 174: 468-474.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42.2: 331-354.

Sapir, Edward. 1916. *Time Perspective in Aboriginal American Culture, a Study in Method*. Geological Survey Memoir 90: No. 13, Anthropological Series. Ottawa: Government Printing Bureau.

Vavilov, Nikolai I. 1926. Studies on the Origin of Cultivated Plants. Leningrad: Institute of Applied Botany and Plant Breeding.