Better Than Chance?
Randomization Models for Evaluating Whether Lexical Similarity Implies Historical Connection

Brett Kessler

Workshop on Alternative Approaches to Language Classification

2007 July 19, 10:00

bkessler@wustl.edu

Key Points

- Languages cannot be assumed to all be related to each other. Proving languages related is a primary task in historical linguistics.
- Widely practiced methodologies for demonstrating language relatedness lack replicability and ways to evaluate the statistical significance of the results.
- We'll discuss techniques for grafting these desiderata onto the traditional comparative method
- Multilateral comparison too can be made into a scientific method for testing whether languages are related to each other.

Probabilistic Approaches

Compute the probability that the observed state of affairs would occur by chance. If it is very low, conclude that the state of affairs cannot be due to chance – the languages must be related.

Collinder's Uralic-Altaic

1947: there is a Uralic-Altaic family, because there are 13 similarities between Uralic and Altaic languages; the odds of this happening by chance are vanishing small.

Hymes

1956: Tlingit and Athapaskan are related, because the odds of having verb prefixes in the same order are 1,216,189,440,000 to 1.

Nichols' 'widow'

1996: Any language that has a word meaning 'widow' that has the consonants *w*, *y*, *dh*, *w*, in that order, must be IE, because the probability is .00000625, which, when multiplied by the number of languages in the world, is less than .05.

Pros and Cons of Probability Approaches

- Yes, improbable similarities point toward similarity
- Most probabilities are incalculable
- Most characters aren't independent, so joint probabilities are not straightforwardly calculable
- Individual low-probability state of affairs are expected by chance

Chance Resemblances

Hawai‘ian *ālike*, English *alike*

Korean *man*, English *man*

Greek *mati* ‘eye’, Malay *mata* ‘eye’

Our Goal

Measure degree to which sound–meaning associations in one language predict sound–meaning associations in another language, while quantifying the probability that any such observed patterns are coincidental: the statistical **significance** of the evidence for language relatedness

The Comparative Method

Gloss	German	Latin
‘heart’	*Herz*	*cord-*
‘horn’	*Horn*	*cornū*
‘dog’	*Hund*	*canis*
‘hundred’	*hundert*	*centum*
‘deer’	*Hirsch*	*cervus*

h : *c* 5 times — a **recurrent sound correspondence**

The Point of the Comparative Method

- The joint occurrence of any two phonemes will be pretty low by chance, since each language has many phonemes
- Iff words are cognate, the chances of recurrent correspondences increase greatly:
  - They start off with perfect correspondences
  - Even if there is change, it is typically **regular**, so we still have correspondences

Ross (1950)

If in Language X .14 of all words start with /s/, and in Language Y .08 of all words start with /ʃ/, then we'd expect $.14 \times .08$ of all concepts to be expressed by a word starting with /s/ in Language X **and** /ʃ/ in Language Y

	GER																	Sum
ENG	f	∅	h	b	v	ʃ	k	z	ʀ	l	n	g	m	t	ts	d	pf	
s	0	0	1	0	0	5	1	6	1	0	0	0	0	0	0	0	0	14
b	1	0	0	5	0	1	1	0	1	0	0	1	0	0	0	0	0	10
h	0	0	6	0	1	0	1	0	0	0	0	0	1	0	0	0	0	9
∅	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
n	0	0	1	0	1	0	1	0	0	0	5	0	0	0	0	0	0	8
f	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
w	1	1	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	7
l	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	5
m	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	5
t	0	0	0	1	0	1	0	0	0	0	0	0	0	0	3	0	0	5
k	0	0	0	0	1	0	3	0	0	0	0	0	0	1	0	0	0	4
ɹ	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	4
d	0	0	1	0	0	1	0	0	0	0	0	0	0	2	0	0	0	4
g	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	3
j	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2
ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Sum	11	9	9	8	8	8	7	7	5	5	5	5	4	3	3	2	1	100

Ross: Pros and Cons

- Great idea to summarize *all* counts in tabular form
- Relevant frequency data comes from the languages themselves, which is possible because sound–meaning associations are arbitrary.
- Plays off the traditional, well-founded assumption that sound change is regular.
- Use of 1st phoneme exploits its relative stability
- No way of computing statistical significance

Villemin (1983)

Pairwise comparisons of Japanese, Korean, and Ainu

Swadesh 200

all and animal ashes at back bad bark because belly big bird bite black blood blow bone breast breathe burn child cloud cold come count cry cut day die dig dirty dog drink dry dull dust ear earth eat egg eye fall far fat father fear feather few fight fire fish five float flow flower fly fog foot four freeze fruit give good grass green guts hair hand he head hear heart heavy here hit hold horn how hunt husband I ice if in kill knife know lake laugh leaf left leg lie live liver long louse man many meat moon mother mountain mouth name narrow near neck new night nose now old one other person play pull push rain red right right river road root rope rotten rub salt sand scratch sea see seed sew sharp short sing sit skin sky sleep small smell smoke smooth snake snow some spear spit split squeeze stab stand star stick stone straight suck sun swell swim tail that there they thick thin think this three throw tie tongue tooth tree true turn two vomit walk warm wash water we wet what when where white who wide wife wind wing wipe with woman woods work worm ye year yellow

Swadesh List: Pros and Cons

- Fixed concept list protects against experimenter bias
- Swadesh concepts are meant to be basic, therefore stable
- Beats going through the whole dictionary
- Not *absolutely* basic and stable
- Stigma of glottochronology

Ringe (1992)

Swadesh 100 (1955)

Significance test: many binomial tests

	German		Sum
English	/ʃ/	Not /ʃ/	
/s/	5	9	14
Not /s/	3	83	86
Sum	8	92	100

/s/ $\sim$ /ʃ/ expectation = $\frac{8 \times 14}{100} = 1.12$, or probability .0112 per word.

Probability of 5 or more events of probability .0112 happening in 100 trials: .005510

Binomial Correction

	GER																	Sum
ENG	f	∅	h	b	v	ʃ	k	z	ʀ	l	n	g	m	t	ts	d	pf	
s	0	0	1	0	0	5	1	6	1	0	0	0	0	0	0	0	0	14
b	1	0	0	5	0	1	1	0	1	0	0	1	0	0	0	0	0	10
h	0	0	6	0	1	0	1	0	0	0	0	0	1	0	0	0	0	9
∅	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
n	0	0	1	0	1	0	1	0	0	0	5	0	0	0	0	0	0	8
f	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
w	1	1	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	7
l	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	5
m	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	5
t	0	0	0	1	0	1	0	0	0	0	0	0	0	0	3	0	0	5
k	0	0	0	0	1	0	3	0	0	0	0	0	0	1	0	0	0	4
ɹ	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	4
d	0	0	1	0	0	1	0	0	0	0	0	0	0	2	0	0	0	4
g	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	3
j	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2
ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Sum	11	9	9	8	8	8	7	7	5	5	5	5	4	3	3	2	1	100

Probability of 16 or more events of $p = .01$ in 48 trials: very small

Empirical Disadvantage of Binomial Correction

Table	Signif. cells	Needed	Connected?
English–German	16	8	yes
English–Latin	7	8	no
—— 200 words	7	9	no
English–French	3	8	no
Albanian–French	3	10	no

Theoretical Disadvantage of Binomials

- Expected frequencies in each cell aren't independent
- Bad approximation when expect much less than 1 entry per cell
- Baxter & Manaster Ramer (1996): hypergeometric is called for

Theoretical Disadvantages of Chi-Squared Tests

$$\sum \frac{(E-O)^2}{E}$$

- Based on difference of each cell from expectation, much broader than linguists' emphasis on recurrences
- Basic requirements of χ^2 distribution don't apply – expected frequencies are far too low – so the normal way of testing significance of χ^2 statistic can't be applied

Permutation Tests

Compare the observed χ^2 statistic to what we get if English and German words (viz. their initial phonemes) are not matched by meaning

There are many ways to match words up if we ignore meaning: try them all

What proportion of those rearrangements (p) has a χ^2 statistic at least as high as the one we get when we match words by meaning?

Permutation Tests: Pros and Cons

- Test gets exactly at what statisticians mean by statistical significance
- Easy to understand
- Flexible
- Universe will die

Monte Carlo Tests

Just like permutation tests, but take a large random sample of possible arrangements of the words

Validation of Monte Carlo Test

Languages		p	Judge
English	German	.001	.610
French	Latin	.001	.591
English	Latin	.001	.318
English	French	.013	.305
German	Latin	.001	.299
Albanian	Latin	.001	.259
French	German	.001	.256
Albanian	French	.003	.216
Albanian	German	.114	.153
Albanian	English	.084	.126

Languages		p
English	Hawai‘ian	.177
Albanian	Hawai‘ian	.186
Albanian	Navajo	.341
Albanian	Turkish	.868
English	Navajo	.073
English	Turkish	.144
French	Hawai‘ian	.262
French	Navajo	.377
French	Turkish	.358
German	Hawai‘ian	.777
German	Navajo	.634
German	Turkish	.984
Hawai‘ian	Latin	.618
Hawai‘ian	Navajo	.233
Hawai‘ian	Turkish	.785
Latin	Navajo	.312
Latin	Turkish	.552
Navajo	Turkish	.224

Monte Carlo Algorithm

1. Build a table counting how many times specific pairs of phonemes co-occur.

2. Compute the χ^2 statistic for that table.

3. Initialize a tally variable t to 1.

4. 9,999 times, shuffle the words in one language. Build the table and compute the χ^2 statistic. If it is at least as high as the real χ^2 statistic, increment t

5. $p = t/10,000$

Variations

- Explore using more natural metrics than χ^2
- Look at more parts of the word, not just the initial consonant
- Tweak the concept list

R^2

Recurrence statistic instead of χ^2 statistic:

Sum over each cell: if count $c \leq 1$, 0; else $(c-1)^2$

R^2 Examples

Languages	p	R^2
English–German	.0000	287
French–Latin	.0000	285
German–Latin	.0000	99
English–Latin	.0005	80
English–French	.0009	49
French–German	.0008	44
Albanian–French	.0005	276
Albanian–Latin	.0031	30
Albanian–German	.2312	24
Albanian–English	.2896	18

Languages	p	R^2
English–Hawai‘ian	.2731	44
Albanian–Hawai‘ian	.0147	46
Albanian–Navajo	.8250	7
Albanian–Turkish	.6366	21
English–Navajo	.8996	14
English–Turkish	.0001	104
French–Hawai‘ian	.4296	38
French–Navajo	.4306	14
French–Turkish	.1131	42
German–Hawai‘ian	.2205	64
German–Navajo	.8586	12
German–Turkish	.8738	31
Hawai‘ian–Latin	.7578	44
Hawai‘ian–Navajo	.5320	35
Hawai‘ian–Turkish	.1754	99
Latin–Navajo	.0561	29
Latin–Turkish	.9370	34
Navajo—Turkish	.3476	30

Beyond P_1?

Why not look at other phonemes besides just the first?

Ringe: run separate tests on various syllable positions

My results with English : Latin:

Position	p
Initial consonant	.0001
2nd consonant of initial cluster	.0676
1st vowel	.1781
1st consonant after 1st vowel	.1014
2nd consonant of 1st cluster after 1st vowel	.0833

How to interpret 5 different test results?

Single Test on All 5 Positions

In each rearrangement, build a frequency-count table for each position; sum R^2 across all tables.

English–Latin: $p = .0131$.

Considering Multiple Positions: Pros and Cons

- Aphaeresis insurance
- In almost all other cases, significance values will degrade considerably, since positions other than the initial are almost always less distinguishable from chance
- Usually is hard to get good clean CCVCC

Tweaking the Concept List

It's OK to tweak the Swadesh list if you can convince others you do so without bias. E.g.:

- Adjust the number of concepts for fun or profit
- You want to avoid words that will spuriously bias the test in one direction or the other

Adjusting the Number of Concepts

There’s nothing sacrosanct about 100 concepts

Size Does Matter

Languages		50	100	Judge
		p	p	
English	German	.000	.000	.610
French	Latin	.000	.000	.591
English	Latin	.049	.005	.318
English	French	.168	.042	.305
German	Latin	.220	.006	.299
Albanian	Latin	.135	.008	.259
French	German	.204	.014	.256
Albanian	French	.051	.003	.216
Albanian	German	.295	.159	.153
Albanian	English	.485	.111	.126

Cognate Proportions in Swadesh Lists

Languages		Swadesh 100	Swadesh 200
Albanian	English	.183	.131
Albanian	French	.217	.207
Albanian	German	.203	.142
Albanian	Latin	.206	.202
English	French	.294	.287
English	German	.768	.686
English	Latin	.347	.294
French	German	.342	.287
French	Latin	.730	.626
German	Latin	.416	.300

Research on Word Stability

Swadesh (1955)

Oswalt (1971)

Kruskal, Dyen, and Black (1973)

Lohr (1999)

Really Short Lists

Lohr/McMahon high-stability words: *four, foot, sun, day, five, new, stand, tooth, name, give, other, eat, I, night, star, wind, three, long, sleep, not, ear, one, thou, two, salt, come, thin, mother, spit, tongue*

Dolgopolsky: *die, eye, heart, I, louse, name, not, tongue, tooth, two, water, what, who*

Lessons

- the more words, the better
- the better words, the better

Avoiding Pernicious Vocabulary

Words that tend to encourage false positives:

- Nonarbitrary vocabulary
- Loanwords
- Language-internal cognates

Nonarbitrary Vocabulary

These tests rely on idea that the only reason for a lot of crosslinguistic similarity between words of matched meaning can be historical contingency, because of the arbitrariness doctrine.

So if some other principle brings about a violation of arbitrariness, we're in trouble; we must expunge the offending words.

Cows Go Moo

Onomatopoeia: lots of languages have cows go /mu/ and cats /miau/.

If we had those calls in our concept list we would be very likely to get a recurring /m/ which would falsely be taken as evidence linking the languages historically.

Possible Offenders on Swadesh List

- ‘bird’ (Swadesh 100). Navajo *tsídii*, which refers to smaller birds. From onomatopoeia *tsíd*, a chirping sound.
- ‘suck’ (Swadesh 200). Albanian *thith*. English *suck*, Proto-Athabascan root /*tˀutˀ/.
- ‘mother’ (Swadesh 200). PIE */ma/
- ‘dirty’ (Swadesh 200). Navajo *baa’ih*
- ‘small’ (Swadesh 100). French *petit*. Proto-Polynesian */ʔiti/

Divergence vs Convergence

These tests show historical connection, which is often precise enough (e.g., any connection at all between Amerind and Asian languages?)

But if you want to specifically test for common parentage (genetic relatedness), you need to expunge loans (e.g., are Japanese and Korean related?)

Loan Counts

Language	S100	Only S200
Albanian	15	25
English	11	19
French	9	15
German	2	4
Latin	0	0
Navajo	0	0
Turkish	5	13

Language-Internal Cognates

The phonetic form of a word isn't arbitrary if it is based on another word in the same list: words that have similar senses tend to have similar forms. That spells trouble.

Identical Albanian Words for Different Concepts

ai	‘he’	‘that’
burrë	‘man’	‘husband’
grua	‘woman’	‘wife’
në	‘at’	‘in’

Identical Hawai‘ian Words for Different Concepts

‘ike	‘see’	‘know’
hele	‘go’	‘come’
‘ili	‘bark’	‘skin’
hua	‘fruit’	‘egg’
kāne	‘man’	‘husband’
lā	‘sun’	‘day’
lā‘au	‘tree’	‘stick’
lepo	‘earth’	‘dirty’
moe	‘sleep’	‘lie’
nui	‘big’	‘many’
wahine	‘woman’	‘wife’

Repeated Roots Lead to Spurious Recurrences

Gloss	Albanian	Hawai‘ian
‘husband’	*burrë*	*kāne*
‘man’	*burrë*	*kāne*
‘wife’	*grua*	*wahine*
‘woman’	*grua*	*wahine*

/b/:/k/ and /g/:/w/ both appear to be recurrent matches

Associations of Sounds to Word Class

Meaning	Class	Lang X	Lang Y
‘see’	verb	/kata/	/wulu/
‘drink’	verb	/kiminu/	/wasup/
‘live’	verb	/kavun/	/wonk/
‘dog’	noun	/spot/	/rovr/
‘hand’	noun	/jad/	/manu/

Greater-than-chance correlation between phonetics and some universal property such as word class can lead to recurrent correspondences.

French *couvrir* ‘cover’, *coudre* ‘sew’, *compter* ‘count’

Multilateral Comparison at a Glance

Engl	Dutch	Germ	Swed	Fren	Port	Span	Welsh	Corn	Swah
one	een	eins	en	un	um	uno	un	un	moja
two	twee	zwei	två	deux	dois	dos	dau	deu	wili
three	drie	drei	tre	trois	três	tres	tri	try	tatu
four	vier	vier	fyra	quatre	quatro	cuatro	pedwar	peswar	ne
five	vijf	fünf	fem	cinq	cinco	cinco	pump	pymp	tano
six	zes	sechs	sex	six	seis	seis	chwech	whegh	sita
seven	zeven	sieben	sju	sept	sete	siete	saith	seyth	saba
eight	acht	acht	åtta	huit	oito	ocho	wyth	eth	nane
nine	negen	neun	nio	neuf	nove	nueve	naw	naw	kenda
ten	tien	zehn	tio	dix	dez	diez	deg	dek	kumi

Your Answer

Engl	Dutch	Germ	Swed	Fren	Port	Span	Welsh	Corn	Swah
one	een	eins	en	un	um	uno	un	un	moja
two	twee	zwei	två	deux	dois	dos	dau	deu	wili
three	drie	drei	tre	trois	três	tres	tri	try	tatu
four	vier	vier	fyra	quatre	quatro	cuatro	pedwar	peswar	ne
five	vijf	fünf	fem	cinq	cinco	cinco	pump	pymp	tano
six	zes	sechs	sex	six	seis	seis	chwech	whegh	sita
seven	zeven	sieben	sju	sept	sete	siete	saith	seyth	saba
eight	acht	acht	åtta	huit	oito	ocho	wyth	eth	nane
nine	negen	neun	nio	neuf	nove	nueve	naw	naw	kenda
ten	tien	zehn	tio	dix	dez	diez	deg	dek	kumi
	Germanic				Romance		Brythonic		

Multilateral Comparison: Pros and Cons

- Based on principle of gradualness of sound change
- It's hard for people to cope with all the data
- Statistical procedures love lots of data

Key Features of Multilateral Comparison

1. Construction of word lists forming tableaux of words for the same concept across languages

2. Use of similarity criteria rather than recurrent sound correspondences

3. Use of a flexible list of concepts

4. Use of multiple words for the same concept in a single language ('burn' is PIE *as-* or *dheg^{w}h-*)

5. Simultaneous comparison across **many** languages

Oswalt 1970, 1998

Compare the first consonants of the words. Consonants are in the same equivalence group if they have the same point of articulation and agree in voicing, stoppage, and nasality, or two out of the three.

Words for the same concept in different languages are similar (= 1) if their initial consonant is in the same equivalence group, else dissimilar (= 0).

Shift test: move words in 2nd column down by 1 and recompute global distance; repeat.

Baxter & Manaster Ramer, 2000

1. Rearrange second column *at random*
2. Compute new distance value
3. See if new value is at least as low as the distance measure for the true, original arrangement of the data
4. Repeat at least 1,000 times.
5. Tally proportion of rearrangements that have a distance value at least as low as the true, original arrangement of the data

Are We There Yet?

1. 🙂 Construction of word lists forming tableaux of words for the same concept across languages

2. 🙂 Use of similarity criteria rather than recurrent sound correspondences

3. 😕 Use of a flexible list of concepts

4. 😕 Use of multiple words for the same concept in a single language (‘burn’ is PIE *as-* or *dheg^{w}h-*)

5. 😕 Simultaneous comparison across **many** languages

Quantifying Phonetic Distance

Given the first consonants of the two words, how far apart is their place of articulation?

Place	Examples	Numeric
labial	/p/, /f/, /m/	0
anterior	/t/, /θ/, /s/, /ʃ/	4
palatal	/c/, /j/	6
velar	/k/, /g/, /ŋ/	9
postvelar	/q/, /h/	10

E.g., English *good* and Finnish *hyvä* have a phonetic dissimilarity of $|9 - 10| = 1$.

Flexible Concept Lists

Start with Swadesh 200

Discard function words up front: *and at because few he here how I if in not some there they this we what when where who with ye*

Attaching Retention Rates

Data: Swadesh (1955); Oswalt (1971); Kruskal, Dyen, and Black (1973) (convert replacement rates by $(1/e^r)$

Decisions: Swadesh (1955: Swadesh 100 list), O'Grady, Black, and Hale (Alpher & Nash 1999), Yakhontov, Dolgopolsky (1986)

Aggregation:

oz = (OGrady + Black + Hale) / 3

booleans = (S100 + S200 + Yakhontov + Dolgopolsky + oz) / 5

aggregate = (booleans + Dyen + KruskalPh + KruskalCu + Swadesh + Oswalt) / 6

Example: 'dirty' = .084; 'eye' = .882

Suitability Scores in Individual Languages

Assign low suitability scores to translations that are:

- missing (e.g., Gothic ‘swim’)
- non-arbitrary (e.g., Latin *māter* ‘mother’)
- loanwords (e.g., Latin *petra*, from Greek)
- repeat a root found more typically elsewhere in the list (e.g., ‘dig’ in Latin = ‘stab’, *fodere*)
- are derived from other forms, e.g., for ‘guts’, Latin *intestīna*, derived from *in*.

Sample Goodness Ratings for Swadesh Concepts

Concept	Goodness
‘two’	0.423006650
‘name’	0.144501558
‘eye’	0.134457565
‘one’	0.018304324
‘three’	0.010189144
‘four’	0.007803799
‘ear’	0.005581617
‘water’	0.002096134
‘horn’	0.001968728
‘drink’	0.001661923
‘new’	0.000841952
‘star’	0.000426747
‘sun’	0.000343778
‘stone’	0.000262479
‘hand’	0.000190347
‘knee’	0.000170032
‘blood’	0.000085625
‘heart’	0.000081208
‘long’	0.000074508
‘night’	0.000049292
‘eat’	0.000045787
‘head’	0.000044570
‘sew’	0.000026580
‘fire’	0.000026497
‘dry’	0.000013498
‘worm’	0.000008007
‘live’	0.000005576
‘person’	0.000005251
‘meat’	0.000003068
‘see’	0.000002684
‘road’	0.000001782
‘black’	0.000000775
‘sleep’	0.000000497
‘cry’	0.000000327
‘big’	0.000000317
‘bite’	0.000000277
‘white’	0.000000241
‘skin’	0.000000200

Multiple Translations

‘Eat’: Latin *edō* vs. Greek *esthíō, édō, éphagon*:

Take the average of all pairwise matchings.

Latin	Greek	Phonetic distance
edō	*esthíō*	0.5
edō	*edō*	0.0
edō	*éphagon*	4.5
AVERAGE		1.7

Multiple Languages

Indo-European		Uralic		Score
Latin	English	Finnish	Nenets	
edō	*etan*	*syödä*	*ŋamč*	3.0

edō	*syödä*	0.5
edō	*ŋamč*	5.5
etan	*syödä*	0.5
etan	*ŋamč*	5.5
AVERAGE		3.0

Multilateral Comparison

1. Gather translations for all the Swadesh words in all the languages
2. Group the languages into two families
3. For each concept, compute the dissimilarity between the families
4. Sum the distance measures across all the concepts

Monte Carlo Significance Testing

1. Measure actual distance d between Indo-European and Uralic.
   Initialize $t = 1$

2. Repeat 999 times:

   (a) Randomly reassign all words to arbitrary meanings.
   (b) Measure randomized distance r between IE and Uralic.
   (c) If $r \leq d$, increment tally t

3. Report significance measure $p = t/1000$.

4. If significant, report strength measure $s = (\bar{r} - d)/\bar{r}$

Randomization Example

Indo-European		Uralic		Gloss
Old English	Latin	Finnish	Nenets	
micel	*magnus*	*iso*	*ŋaarka*	‘big’
blōd	*sanguis*	*veri*	*sielw*	‘blood’
hund	*canis*	*koira*	*wõneko*	‘dog’

Keep together words in the same language family:

Indo-European		Uralic	
Old English	Latin	Finnish	Nenets
micel	*magnus*	*veri*	*sielw*
blōd	*sanguis*	*iso*	*ŋaarka*
hund	*canis*	*koira*	*wõneko*

The Languages Used

- HAW (Hawai‘ian) $\in$ Austronesian
- FIN (Finnish) and CHM (Mari) $\in$ Finno-Baltic $\in$ Finno-Ugric $\in$ Uralic
- HUN (Hungarian) $\in$ Finno-Ugric $\in$ Uralic
- YRK (Nenets) $\in$ Uralic
- ANG (Old English), GOH (Old High German), NON (Old Norse) and GOT (Gothic) $\in$ Germanic $\in$ IE
- CHU (Old Church Slavonic) and LIT (Lithuanian) $\in$ Balto-Slavic $\in$ IE
- SGA (Old Irish), LAT (Latin), GRC (Classical Greek), SAN (Sanskrit), and ALB (Albanian) $\in$ IE

Nearest-Neighbor Clustering

1. Do bilateral comparison for each pair of languages. Of the pairs with $p \leq .05$, store their strength s. Stop if no pair has $p \leq .05$.

2. Take the pair with highest s and merge them into one group, treating it henceforth as a single language.

3. Go to step 1.

E.g., the pair English–German is significant at $p < .0001$, and $s = 88\%$, larger than all other s. So in the next round we treat the group English–German as a single merged language where, e.g., *dēor* and *tior* are treated as alternative words for 'animal'.

C1-Place

p	dist	s		Languages grouped
.001	14.5	.88	ANG	GOH
.001	23.8	.81	GOT	ANG+GOH
.001	34.3	.74	NON	GOT+ANG+GOH
.001	63.0	.52	CHU	LIT
.002	74.0	.43	FIN	HUN
.001	71.7	.40	LAT	NON+GOT+ANG+GOH
.001	76.3	.38	ALB	CHU+LIT
.001	91.0	.35	CHM	FIN+HUN
.003	91.3	.26	SGA	ALB+CHU+LIT
.008	98.3	.21	SAN	LAT+NON+GOT+ANG+GOH
.027	118.1	.15	GRK	SAN+LAT+NON+GOT+ANG+GOH
.016	112.7	.11	SGA+ALB+CHU+LIT	GRK+SAN+LAT+NON+GOT+ANG+GOH

Dolgopolsky metrics

All sounds are placed into one of 10 equivalency classes. Two sounds in the same class are distant 0, else 1

- P - labial obstruents
- T - dental or apical obstruents
- S - sibilants
- K - palatals, dorsals, and postalveolar affricates
- M - /m/
- N - other nasals
- R - liquids
- W - rounded semivowels
- J - /j/
- O - vowels, dorsal nasals, and glottals

P1-Dolgopolsky

p	dist	s		Languages grouped
.001	5.1	.84	ANG	GOH
.001	5.7	.82	GOT	ANG+GOH
.001	8.0	.75	NON	GOT+ANG+GOH
.001	15.5	.52	CHU	LIT
.001	20.4	.38	LAT	NON+GOT+ANG+GOH
.001	21.0	.36	FIN	CHM
.001	21.5	.33	SGA	ALB
.001	22.0	.33	HUN	FIN+CHM
.001	22.5	.30	SAN	LAT+NON+GOT+ANG+GOH
.001	25.6	.23	CHU+LIT	SAN+LAT+NON+GOT+ANG+GOH
.001	27.0	.18	SGA+ALB	CHU+LIT+SAN+LAT+NON+GOT+ANG+GOH
.001	26.8	.16	GRK	SGA+ALB+CHU+LIT+SAN+LAT+NON+GOT+ANG+GOH
.008	29.0	.11	YRK	HUN+FIN+CHM

Grimes Metrics

Each phoneme is a 6D entity. Distance between sounds is the sum of the distance on all 6 dimensions.

Each dimension takes on one integer value out of from 2 to 7 possibilities. Values are assigned ordinally. E.g.:

Point of articulation:

1 bilabial
2 labiodental
3 interdental
4 apical
5 laminal
6 dorsal
7 nonbuccal

Grimes Metric Example

Feature	/tʷ/	/iː/	|Diff|
Place	4 apical	4 front	0
Aperture	1 stop	4 high vowel	3
Length	3 normal	4 long	1
Secondary	1 vowel-shaped	0 normal	1
Nasality	0 oral	0 oral	0
Glottal aperture	2 open	1 voiced	1
Sum			6

Grimes

p	dist	s	Languages grouped	
.001	32.3	.82	ANG	GOT
.001	50.0	.73	GOH	ANG+GOT
.001	62.6	.66	NON	GOH+ANG+GOT
.001	100.5	.41	CHU	LIT
.001	119.5	.39	FIN	CHM
.001	132.4	.36	LAT	NON+GOH+ANG+GOT
.001	127.8	0.35	CHM	FIN+HUN
.001	129.0	0.34	SAN	LAT+NON+GOH+ANG+GOT
.001	117.7	0.27	ALB	CHU+LIT
.001	148.2	0.19	SGA	ALB+CHU+LIT
.001	158.4	0.18	GRK	SGA+ALB+CHU+LIT
.001	164.2	0.16	SAN+LAT+NON+GOH+ANG+GOT	GRK+SGA+ALB+CHU+LIT
.002	160.2	0.12	YRK	CHM+FIN+HUN

C1-Voice

p	dist	s	Languages grouped	
.001	2.8	.85	ANG	NON
.001	3.3	.82	GOT	ANG+NON
.001	5.4	.71	GOH	GOT+ANG+NON
.001	6.5	.66	CHU	LIT
.001	8.7	.52	HUN	CHM
.001	9.5	.50	YRK	CHU+LIT
.001	9.3	.48	FIN	HUN+CHM
.002	10.7	.44	LAT	GRK
.008	11.9	.35	SAN	HAW
.009	13.7	.28	SGA	LAT+GRK
.010	13.4	.28	GOH+GOT+ANG+NON	SAN+HAW
.014	14.8	.22	ALB	SGA+LAT+GRK
.045	15.5	.18	YRK+CHU+LIT	GOH+GOT+ANG+NON+SAN+HAW

What??

1. Maybe these languages are all related
2. Maybe there are undetected loans
3. Maybe just a stochastic fluke (or Bonferroni was right)
4. Maybe choice of initial consonant is not purely arbitrary
5. Maybe initial consonant correlates with something that is not purely arbitrary

Conclusions I

Results so far seem broadly comparable to what we get by applying the traditional comparative method.

This broad equity, combined with the new benefits of explicit quantification and inferential statistics, makes them worth exploring further.

Results suggest that some of the very bold claims made by multilateralists may be due to failure to apply sufficient statistical checks; but the basic idea of multilateral comparison is still valid, when done properly.

Conclusions II

There's not yet a clear winner between recurrent sound correspondences and the phonetic approach, though the former has more interesting spin-offs:

English–Latin correspondences at $p \leq .01$:

f$\sim$p, h$\sim$k, l$\sim$j (!), n$\sim$n, s$\sim$s, t$\sim$t

English–Latin Pairs at $p \leq .01$

not	*non*
horn	*cornu*
night	*nox*
hot	*calidus*
heart	*cor*
stands	*stat*
name	*nomen*
new	*novus*
sits	*sedet*
star	*stella*
foot	*pes*

Conclusions III

- Best phonetic comparison metrics seem to be those that compare initial phonemes, especially on place of articulation. It can be dangerous and unproductive to get any more complicated.
- Swadesh 200 is no better than Swadesh 100, but best bet may be to whittle down the list based on actual research
- Monte Carlo is not a town in Monaco