GENETIC RELATIONSHIP AND CHANCE: A NONMATHEMATICIAN'S APPROACH

John E. McLaughlin

Utah State University

DON'T PANIC! These words, written in large friendly letters[1], should allay your fears about my approach to the issues of chance, probability, randomness, and genetic relationship in languages. I'm not a mathematician and it takes a good night's sleep, the wife out of the house for the day, and a comfortable recliner in order to ferret out any meaning at all in the complex formulas presented in most papers on this subject. In this way, I probably reflect the knowledge and interest of most linguists doing comparative and historical linguistics. But this issue is not one that can be ignored despite our difficulty in tackling the math involved. It is the subject of a growing body of writing and is beginning to form one of the most critical elements of the debate over Nostratic, Greenberg's Amerind, Ruhlen's Proto-World, and even the basic question of how far back in time can we demonstrate a genetic relationship.

As I began reading this body of literature, I quickly realized that there was a critical debate developing over the basic mathematical assumptions. The debate has hinged on two issues: 1) Is there really a difference in the rates of retention between the so-called "basic vocabulary" and the rest of the lexicon; and 2) are multilateral sets less likely to be affected by chance resemblances or not? It is the latter issue which sparked my curiosity. The problem with the debates over this matter is that to demonstrate the validity of any of the hypotheses, pairs of real languages were used. The first problem is that one can never categorically rule out a relationship. The absence of a proven relationship does not automatically prove that the two languages cannot be related. There are also a multitude of areal features to consider even between unrelated languages. Therefore there is always an unknown element when using real languages to demonstrate the factor of chance. The second problem is the element of semantic content. Can one legitimately compare 'daughter' with 'girl', or 'be' with 'become', or 'door' with 'entry'? Most comparative linguists, working outside the realm of mathematics, have no real problem with comparing any of these obviously very closely related concepts. A great many words in any dictionary contain multiple meanings for each item. Which one do we choose as the primary one for comparison? Yet each additional semantic possibility or semantic ambiguity that we add to a possible comparative group increases the chance of random matches.

With these problems in mind, I decided to approach the issue of demonstrating a factor of chance from a different perspective. Instead of using real languages, which are always subject to the possibility (no matter how remote) of actual relationship or pseudo-relationship due to areal similarities, I designed a computer program in Visual Basic 5.0 to produce a random lexicon for ten languages and, using strict rules of correspondence between sounds, had the computer find all the

---

[1]Written on the cover of the fictional *Hitchhiker's Guide to the Galaxy*, described by Douglas Adams in his book of the same name

binary pairs in these languages that would count as cognates to a typical comparative linguist. Since these computer-generated (CG) languages had no possibility of genetic relationship, and since the semantic content could be precisely controlled at various levels, it provided very reliable information about the chances of random cognate sets between unrelated languages. In essence, it provides an experimentally derived basis of comparison rather than a mathematically-derived one.

In designing the ten CG languages, I divided the set of languages into four groups–three languages with small consonant inventories (less than 20), three languages with medium-sized consonant inventories (20-30), two languages with large inventories (30-40), and two languages with very large inventories (over 40). I based the phonologies of these CG languages on real-world languages, using both the actual phonemic inventories and the frequency of occurrence for each of the phonemes. In addition, as a control measure, two of the small CG languages and both of the very large CG languages were based on the same two real-world languages. The two identical small languages were based on Shoshone; the other small language on Zuni. The three medium-sized languages were based on reconstructed Proto-Indo-European, Hungarian, and English. The two large languages were based on Eastern Keres and Lushootseed, and the two very large languages were based on Heiltsuk. Except for English and Proto-Indo-European, all these languages represent unrelated language families, and each has a different consonant and vowel inventory. Table 1 illustrates the inventories for each of the eight real-world languages and matches them to the ten CG languages.

Table 1. Eight Real World Languages Were Used as the Basis for the Ten CG Languages

CG#	**Language (family): consonants; vowels**
L1/2	Shoshone (Uto-Aztecan): p, t, ts, k, kw, s, h, m, n, w, j; i, e, a, ɨ, o, u
L3	Zuni (isolate): p, t, ts, tʃ, k, kw, s, ʃ, h, m, n, l, ɬ, w, j; i, e, a, o, u
L4	Indo-European (reconstructed): bh, b, p, dh, d, t, gh, g, k, gwh, gw, kw, s, h, m, n, l, r, w, j; i, e, a, o, u
L5	Hungarian (Uralic): b, p, d, t, ts, ɟ, c, tʃ, g, k, v, f, z, s, ʒ, ʃ, h, m, n, ɲ, l, r, j; i, e, a, o, œ, u, y
L6	English (Indo-European): b, p, dʒ, t, d, tʃ, g, k, v, f, ð, θ, z, s, ʒ, ʃ, h, m, n, ŋ, l, r, w, j; i, ɪ, e, ɛ, æ, a, ə, o, ɔ, u, ʊ
L7	Eastern Keres (Keresan): b, p, p', d, t, t', ts, ts', tʂ, tʂ', ɟ, tʃ, tʃ', g, k, k', z, s, s', ʐ, ʂ, ʂ', ʃ, ʃ', h, m, m', n, n', r, w, w', j, j'; i, e, a, ə, u
L8	Lushootseed (Salishan): b, p, p', d, t, t', dz, ts, ts', dʒ, tʃ, tʃ', g, k, k', gw, kw, k'w, q, q', qw, q'w, s, ʃ, xw, χ, χw, h, m, n, tl', l, ɬ, w, j; i, a, ə, u
L9/10	Heiltsuk (Wakashan): b, p, p', d, t, t', dz, ts, ts', g, k, k', gw, kw, k'w, ɢ, q, q', ɢw, qw, q'w, s, x, xw, χ, χw, h, h', m, m', n, n', dl, tl, tl', l, ɬ, l', w, w', j, j'; i, a, u

The program constructed a random vocabulary of 1,000 words for each of the ten languages. Each of these words consisted of a CVC sequence. I chose a CVC sequence since that tends to be the most commonly used sequence in comparisons and led to a two-tiered comparison of the forms by the computer. First, do the two consonants match, and second, does the vowel also match?

I then constructed a table of correspondences to use in comparing the disparate phonologies to one another. I basically made sure, using commonly found correspondences, that each sound in each language was part of a regular correspondence set. Thus, in the languages without glottalized consonants, for example, the plain versions matched both the plain versions and the glottalized versions in the languages that have them. The same basic principles were used for correlating the matches between uvulars and velars, lateral and rhotic approximants, fricatives, etc. Table 2 shows the Table of Correspondences.

Table 2. *The Table of Correspondences Guided the Process of Matching Forms*

L1-2	L3	L4	L5	L6	L7	L8	L9-10
p	p	bh	b	b	b	b	b
p	p	b	b	b	b	b	b
p	p	p	p	p	p	p	p
p	p	p	p	p’	p’	p’	p’
t	t	dh	d	d	d	d	d
t	t	d	d	d	d	d	d
t	t	t	t	t	t	t	t
t	t	t	t	t	t’	t’	t’
ts	ts	d	ts	dʒ	ts	dz	dz
ts	ts	t	ts	tʃ	ts	ts	ts
ts	ts	t	ts	tʃ	ts’	ts’	ts’
ts	ts	t	ts	tʃ	tʂ	ts	ts
ts	ts	t	ts	tʃ	tʂ’	ts’	ts’
t	t	d	ɟ	d	ɟ	d	d
t	t	t	c	t	t	t	t
ts	tʃ	d	tʃ	dʒ	tʃ	dʒ	dz
ts	tʃ	t	tʃ	tʃ	tʃ	tʃ	ts
ts	tʃ	t	tʃ	tʃ	tʃ’	tʃ’	ts’
k	k	gh	g	g	g	g	g
k	k	g	g	g	g	g	g
k	k	k	k	k	k	k	k
k	k	k	k	k	k’	k’	k’
kw	kw	gwh	g	g	g	gw	gw
kw	kw	gw	g	g	g	gw	gw
kw	kw	kw	k	k	k	kw	kw
kw	kw	kw	k	k	k’	k’w	k’w
k	k	g	g	g	g	g	ɢ
k	k	k	k	k	k	q	q
k	k	k	k	k	k’	q’	q’
kw	kw	gw	g	g	g	gw	ɢw
kw	kw	kw	k	k	k	qw	qw
kw	kw	kw	k	k	k’	q’w	q’w
p	p	b	v	v	b	b	b
p	p	p	f	f	p	p	p
t	t	d	d	ð	d	d	d

t	t	t	t	θ	t	t	t
s	s	s	z	z	z	s	s
s	s	s	s	s	s	s	s
s	s	s	s	s	s'	s	s
s	s	s	z	z	ʐ	s	s
s	s	s	s	s	ʂ	s	s
s	s	s	s	s	ʂ'	s	s
s	ʃ	s	ʒ	ʒ	ʃ	ʃ	s
s	ʃ	s	ʃ	ʃ	ʃ	ʃ	s
s	ʃ	s	ʃ	ʃ	ʃ'	ʃ	s
k	k	k	k	k	k	k	x
kw	kw	kw	k	k	k	xw	xw
k	k	k	k	k	k	χ	χ
kw	kw	kw	k	k	k	χw	χw
h	h	h	h	h	h	h	h
h	h	h	h	h	h	h	h'
m	m	m	m	m	m	m	m
m	m	m	m	m	m'	m	m'
n	n	n	n	n	n	n	n
n	n	n	n	n	n'	n	n'
n	n	n	ɲ	n	n	n	n
n	n	n	n	ŋ	n	n	n
t	t	d	d	d	d	d	dl
t	t	t	t	t	t	t	tl
t	t	t	t	t	t'	tl'	tl'
t	l	l	l	l	r	l	l
t	ɬ	l	l	l	r	ɬ	ɬ
t	l	l	l	l	r	l	l'
t	l	r	r	r	r	l	l
w	w	w	m	w	w	w	w
w	w	w	m	w	w'	w	w'
j	j	j	j	j	j	j	j
j	j	j	j	j	j'	j	j'
i	i	i	i	i	i	i	i
i	i	i	i	ɪ	i	i	i
e	e	e	e	e	e	i	i
e	e	e	e	ɛ	e	i	i
a	a	a	a	æ	a	a	a
a	a	a	a	a	a	a	a
ɨ	a	a	a	ə	ə	ə	a
o	o	o	o	o	u	u	u
o	o	o	o	ɔ	u	u	u
o	o	o	œ	o	u	u	u
u	u	u	u	u	u	u	u
u	u	u	u	ʊ	u	u	u

u u u y u u u u

The final process in the construction of the program was to decide how to deal with semantics. The problem was solved quite simply–each word was numbered as it was generated. Exact semantic matches (as in comparing 'eat' in Language A with 'eat' in Language B) were simply a case of comparing Word 1 in L1 to Word 1 in L2, etc. Dealing with non-exact semantic matches (as in comparing 'girl' with 'daughter') was a more complicated issue. I solved the problem by using a moving vector approach and taking advantage of what we may call the "Thesaurus Effect". The Thesaurus Effect is starting with a word and then moving through the choices in a thesaurus until one arrives at a word which has a completely different, unrelated meaning to the original word. We've all played this game at one time or another and are always amazed at the permutations we can come up with. I used this Thesaurus Effect in the program by comparing Word 1 in L1 with Words 1-10 in L2 for a semantic latitude typical of most long-range comparisons. The program then compared Word 2 in L1 with Words 2-11 in L2, etc. Thus, a semantic latitude of 1 represented extremely tight 'girl' equals 'girl' comparisons, a semantic latitude of 5 represented typical 'girl' equals 'girl', 'child', or 'daughter' comparisons, and a semantic latitude of 10 represented looser 'girl' equals 'girl', 'child', 'daughter', 'sister', 'niece', 'female', 'woman', 'sibling' comparisons.

The program reported several pieces of information:

- The vocabularies of the ten languages in the last iteration
- The pairs of words which were found to be matches based on a two-consonant comparison in the last iteration
- In 100 iterations, the average number of pairs found to match for each pair of languages when comparing just the consonants, and the whole form, and when using a semantic latitude of 1, and a semantic latitude of 10
- In 100 iterations, the minimum and maximum numbers of matching pairs found in any of the iterations for each pair of languages under the same conditions as the averages

The run on which I am basing the following discussion consisted of 100 iterations of generating 1,000 words and using a semantic latitude of 10, both using the Table of Correspondences and not using it (that is, insisting on p matching only p and not matching p'). Please refer to Tables 3 through 6 on the following pages for the discussion that follows. Tables 3 and 4 show the results from requiring exact matches between sounds and Tables 5 and 6 show the results from using the Table of Correspondences. The charts on the left side show the results of matching the two consonants of each form. The charts on the right side show the results of matching all three elements of each word. Tables 3 and 5 show averages; Tables 4 and 6 show the highest number achieved in the 100 iterations.

First, we'll examine the most restrictive of the charts. Look at the top right chart on Table 3. This chart illustrates the results of finding exact matches between all three elements of each word and allowing no semantic variation. This is equivalent to comparing Shoshone kimma 'come' to Panamint kimma 'come'. Two general rules begin to stand out. The first generalization is the greater the difference between the phonological inventories of the two languages, the lower the number of matches found. So L1 and L2, which have identical phonologies, show the greatest number of matches using these restrictive criteria. Notice also that L9 and L10, which also have identical

Strict Matches between Consonants and Vowels (a=a, b=b, etc.) 45,000 possible pairs

Averages of 100 iterations

Numbers of binary sets for semantic range of 1 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total pairs
L1		17	7	5	3	5	2	1	1	1	42	**42**
L2			7	4	3	4	2	1	1	1	23	**40**
L3				3	4	2	2	1	1	1	14	**28**
L4					2	2	2	0	0	0	6	**18**
L5						2	1	1	0	1	5	**17**
L6							1	1	0	0	2	**17**
L7								1	1	1	3	**13**
L8									1	1	2	**8**
L9										2	2	**7**
L10											0	**8**
	0	17	14	12	12	15	10	6	5	8		99
												0.22%

Numbers of binary sets for semantic range of 1 in 1000 words, CV

0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total pairs
L1		3	1	1	0	0	0	0	0	0	5	**5**
L2			1	1	1	0	0	0	0	0	3	**6**
L3				1	1	0	0	0	0	0	2	**4**
L4					1	0	0	0	0	0	1	**4**
L5						0	0	0	0	0	0	**3**
L6							0	0	0	0	0	**0**
L7								0	0	0	0	**0**
L8									0	0	0	**0**
L9										1	1	**1**
L10											0	**1**
	0	3	2	3	3	0	0	0	0	1		**12**
												0.03%

Numbers of binary sets for semantic range of 10 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total pairs
L1		166	70	44	34	46	21	9	9	9	408	**408**
L2			70	44	32	46	21	10	9	9	241	**407**
L3				32	43	25	18	10	10	10	148	**288**
L4					21	20	16	4	4	4	69	**189**
L5						22	12	6	6	6	52	**182**
L6							11	7	4	4	26	**185**
L7								5	8	8	21	**120**
L8									6	6	12	**63**
L9										15	15	**71**
L10											0	**71**
	0	166	140	120	130	159	99	51	56	71		**992**
												2.20%

Numbers of binary sets for semantic range of 10 in 1000 words, CV

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total pairs
L1		30	13	10	6	3	3	2	2	2	71	**71**
L2			14	8	5	3	3	1	1	2	37	**67**
L3				9	8	3	3	3	3	3	32	**59**
L4					4	2	3	1	1	1	12	**39**
L5						2	2	1	1	1	7	**30**
L6							1	1	0	0	2	**15**
L7								1	2	2	5	**20**
L8									2	2	4	**14**
L9										6	6	**18**
L10											0	**19**
	0	30	27	27	23	13	15	10	12	19		**176**
												0.39%

Table 3. Average Number of Binary Matches using Exact Matches

Strict Matches between Consonants and Vowels (a=a, b=b, etc.)

Maximums of 100 iterations

Numbers of binary sets for semantic range of 1 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		30	15	12	11	14	9	4	4	3	102	**102**
L2			17	15	8	11	8	4	4	4	71	**101**
L3				8	10	7	6	7	4	4	46	**78**
L4					7	6	5	2	2	2	24	**59**
L5						7	5	3	2	3	20	**56**
L6							4	3	2	3	12	**57**
L7								3	4	3	10	**47**
L8									3	4	7	**33**
L9										6	6	**31**
L10											0	**32**
	0	30	32	35	36	45	37	26	25	32		**298**
												0.66%

Numbers of binary sets for semantic range of 1 in 1000 words, CVC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		10	5	3	3	2	3	2	4	2	34	**34**
L2			5	3	3	3	3	2	2	1	22	**32**
L3				4	4	2	2	2	2	2	18	**28**
L4					3	2	3	1	1	1	11	**21**
L5						2	2	1	1	1	7	**20**
L6							1	1	1	1	4	**15**
L7								1	2	2	5	**19**
L8									2	3	5	**15**
L9										4	4	**19**
L10											0	**17**
	0	10	10	10	13	11	14	10	15	17		**110**
												0.24%

Numbers of binary sets for semantic range of 10 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		197	99	75	56	74	35	19	18	17	590	**590**
L2			96	61	56	70	33	21	17	17	371	**568**
L3				47	62	44	26	20	17	18	234	**429**
L4					33	32	28	9	11	9	122	**305**
L5						36	20	13	14	15	98	**305**
L6							20	19	9	10	58	**314**
L7								13	16	16	45	**207**
L8									14	14	28	**142**
L9										24	24	**140**
L10											0	**140**
	0	197	195	183	207	256	162	114	116	140		**1,570**
												3.49%

Numbers of binary sets for semantic range of 10 in 1000 words, CVC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		46	26	24	11	10	8	8	9	6	148	**148**
L2			27	13	12	9	8	5	7	6	87	**133**
L3				18	14	8	8	8	7	9	72	**125**
L4					8	6	8	3	4	3	32	**87**
L5						5	7	6	5	5	28	**73**
L6							4	4	2	3	13	**51**
L7								5	7	7	19	**62**
L8									6	6	12	**51**
L9										12	12	**59**
L10											0	**57**
	0	46	53	55	45	38	43	39	47	57		**423**
												0.94%

Table 4. Maximum Numbers of Binary Matches using Exact Matches

Table Matches between Consonants and Vowels (a=a or a', b=b or b', etc.)

Averages of 100 iterations

Numbers of binary sets for semantic range of 1 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		**Total pairs**
L1		17	14	17	19	24	11	14	10	10	136	**136**
L2			13	16	18	24	11	14	10	10	116	**133**
L3				11	10	9	7	8	11	12	68	**95**
L4					8	5	6	4	4	3	30	**74**
L5						3	4	6	8	8	29	**84**
L6							3	4	3	3	13	**78**
L7								2	2	2	6	**48**
L8									1	1	2	**54**
L9										2	2	**51**
L10											0	**51**
	0	17	27	44	55	65	42	52	49	51		**402**
												0.89%

Numbers of binary sets for semantic range of 1 in 1000 words, CVC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		**Total pairs**
L1		3	4	5	4	4	3	4	4	4	35	**35**
L2			4	5	4	4	2	3	3	3	28	**31**
L3				3	2	3	2	4	4	4	22	**30**
L4					2	1	2	1	1	1	8	**21**
L5						1	1	2	2	2	8	**20**
L6							0	1	1	1	3	**16**
L7								0	1	1	2	**12**
L8									0	0	0	**15**
L9										1	1	**17**
L10											0	**17**
	0	3	8	13	12	13	10	15	16	17		**107**
												0.24%

Numbers of binary sets for semantic range of 10 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		**Total pairs**
L1		166	137	167	188	246	108	134	101	100	1,347	**1,347**
L2			138	167	185	247	105	135	101	100	1,178	**1,344**
L3				105	104	89	75	84	117	120	694	**969**
L4					79	49	64	39	32	32	295	**734**
L5						32	36	59	76	75	278	**834**
L6							31	38	32	31	132	**795**
L7								17	22	22	61	**480**
L8									12	12	24	**530**
L9										15	15	**508**
L10											0	**507**
	0	166	275	439	556	663	419	506	493	507		**4,024**
												8.94%

Numbers of binary sets for semantic range of 10 in 1000 words, CVC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		**Total pairs**
L1		30	37	48	40	44	23	35	35	35	327	**327**
L2			38	47	39	44	23	34	35	34	294	**324**
L3				30	20	27	21	34	43	43	218	**293**
L4					18	15	19	15	12	12	91	**216**
L5						7	9	17	23	22	78	**195**
L6							5	9	12	11	37	**174**
L7								4	7	8	19	**119**
L8									5	5	10	**158**
L9										6	6	**178**
L10											0	**176**
	0	30	75	125	117	137	100	148	172	176		**1,080**
												2.40%

Table 5. Average Number of Binary Matches using Table of Correspondences

Table Matches between Consonants and Vowels (a=a or a', b=b or b', etc.)

Maximums of 100 iterations

Numbers of binary sets for semantic range of 1 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		30	24	27	26	39	20	23	19	19	227	**227**
L2			24	26	31	39	26	22	18	24	210	**240**
L3				20	19	17	15	20	19	19	129	**177**
L4					17	10	14	10	10	8	69	**142**
L5						8	7	13	17	15	60	**153**
L6							9	8	7	8	32	**145**
L7								6	6	6	18	**109**
L8									3	5	8	**110**
L9										6	6	**105**
L10											0	**110**
	0	30	48	73	93	113	91	102	99	110		**759**
												1.69%

Numbers of binary sets for semantic range of 1 in 1000 words, CVC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		10	9	12	9	10	7	10	9	11	87	**87**
L2			9	11	10	9	7	9	9	9	73	**83**
L3				8	7	7	6	8	10	13	59	**77**
L4					7	5	6	6	4	6	34	**65**
L5						3	4	6	9	8	30	**63**
L6							4	3	4	4	15	**49**
L7								4	3	3	10	**44**
L8									3	3	6	**52**
L9										4	4	**55**
L10											0	**61**
	0	10	18	31	33	34	34	46	51	61		**318**
												0.71%

Numbers of binary sets for semantic range of 10 in 1000 words, CC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		197	170	222	227	290	133	169	124	123	1,655	**1,655**
L2			172	198	224	303	131	184	139	119	1,470	**1,667**
L3				133	131	117	108	104	145	161	899	**1,241**
L4					98	64	86	59	49	48	404	**957**
L5						46	50	87	103	104	390	**1,070**
L6							43	53	52	57	205	**1,025**
L7								27	35	34	96	**647**
L8									22	19	41	**724**
L9										24	24	**693**
L10											0	**689**
	0	197	342	553	680	820	551	683	669	689		**5,184**
												11.52%

Numbers of binary sets for semantic range of 10 in 1000 words, CVC

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10		Total Pairs
L1		46	54	78	60	69	39	49	48	52	495	**495**
L2			54	72	50	64	34	54	55	54	437	**483**
L3				45	30	41	30	50	59	62	317	**425**
L4					27	27	29	26	23	22	154	**349**
L5						16	17	30	37	35	135	**302**
L6							10	21	19	22	72	**289**
L7								12	14	16	42	**201**
L8									12	10	22	**264**
L9										12	12	**279**
L10											0	**285**
	0	46	108	195	167	217	159	242	267	285		**1,686**
												3.75%

Table 6. Maximum Numbers of Binary Matches using Table of Correspondences

phonologies, show more matches with each other than with any other language they are compared to. The second generalization is that the larger the phonological inventory, the fewer matches will be found. While the identical L1 and L2 have a small phonological inventory and average three matches out of 1,000, the identical L9 and L10 only show 1 match out of 1,000 with their very large phonological inventories.

Compare this chart, which requires exact phonological matches with the top right chart on Table 5, which uses the Table of Correspondences, but otherwise with the same tight restrictions on semantics and matching all three elements of each word. While the number of matches between L1 and L2 and between L9 and L10 remain the same since these pairs do not use the Table of Correspondences, all the other comparisons between languages show more matches. The second of our two generalizations still holds–the languages with smaller inventories have more matches than the languages with larger inventories. Once we begin to use the Table of Correspondences, however, the radical difference between identical phonologies and non-identical phonologies is not as great, although it can still be seen in the charts requiring only a two consonant match.

We've now seen the number of matches in the most restrictive circumstances–12 pairs out of a possible 45,000 for an exact phonological match and 107 pairs out of a possible 45,000 for a Table of Correspondences match, or 0.03% and 0.24%, respectively. Now we turn to the least restrictive circumstances for a match. Look at the second chart on the left side of Table 3. This shows the average number of matches between two consonants allowing a semantic range of 10, but requiring an exact phonological match. Notice how much more the identical languages–L1 and L2 and L9 and L10–stand out in terms of number of matches between them. The number of matches between L1 and L2 is consistently about twice as many as between either of these languages and L3, another small, but non-identical inventory. The same is true for the number of matches between L9 and L10 compared to the number of matches between either of these languages and L8, with a large phonological inventory. Yet the number of matches between L1 and L2 is up to 11 times higher than the number of matches between L9 and L10, thus clearly demonstrating our two generalizations that the chance of random matches increases with similar and smaller phonologies when not using a Table of Correspondences.

Now look at the corresponding chart on Table 5. This chart is probably the most typical of the type of comparison practiced by most comparative linguists, especially those seeking to demonstrate long-range groupings. It recognizes a semantic leeway of 10 and matches just the two consonants on the Table of Correspondences. Notice that the generalization about smaller phonologies still holds true here, with the languages classed as small (fewer than 20 consonants) having matches with other languages five to six times as often as the languages classed as large or very large (30 or more consonants). Now look at the number of matched pairs–4,024 or 8.94% of a possible 45,000. With 1,000 words in each of the the lexicons, this means that there should be an average of four pairs per lexical item. This may be expressed in one of two ways or a combination of the two. The first way that this might show up is in pairs illustrating the same sounds. With four pairs in a lexical item, this may mean an interlocking set of at least three languages showing the same correspondences in each of the forms. The second way that this might show up is in four unrelated pairs of words between eight of the languages. Usually, a combination of the two types of pairings is seen.

Now look at the bottom left chart on Table 6. This is where the maximum values are given for each of the language pairings out of 100 iterations with a semantic leeway of 10 and only matching the two consonants on the Table of Correspondences. Note that the numbers are at least 20% higher than they are in the averages chart we were just looking at. Also note that the number of pairs has risen to 5,184, or an average of five matches for each of the 1,000 lexical items. Obviously, this is quite relevant to the question of how likely it is to find multilateral comparisons based on chance alone.

Table 7 on the following page shows the some of the sets that came up for lexical items 900-1,000 during one run of the program. Rather than showing the individual pairs, I have lumped the related pairs together to form cognate sets that illustrate correspondence sets for each of the two consonants. The first 20 columns show the word and number for the forms in the ten languages. The final two columns show the so-called "proto-consonants" (one for each correspondence set on the Table of Correspondences) and the number of languages represented in each of the cognate sets. This table began as approximately 341 pairs. The full number of pairs for this iteration was 3411, actually 613 pairs less than the average. There is a good deal more collapsing of sets that could be done, but the current chart was done very precisely according to rule. Note that each of the sounds of the "proto-language" are illustrated by multiple cognate sets and in both initial and final position. Looking at this chart as it stands, many linguists would see at least a suggestive start for further research into a genetic relationship.

What happens to the chances of random matches when we loosen the bonds of comparison even more? For example, if we only compared initial consonants then imagine what Tables 5 and 6 would look like (the numbers would more than double). What if we used longer words? I have used CVC as a standard form, but we often find an initial syllable compared to a final syllable and vice versa. What this does to the numbers in the charts in Tables 3 through 6 is to double them.

In summary, I haven't given any rock hard figure or calculation to determine whether a particular comparison exceeds the threshold of chance possibility. Instead, I have found two generalities–the more similar and the smaller the phonological inventory of the languages being compared, the greater the likelihood of random matches. I have also found that multilateral comparison also increases the chance of finding multiple languages showing two consonant correspondences in particular lexical forms, and giving the overall impression of a solid linguistic grouping with a full range of proto-forms.

Table 7. Strictly Defined "Cognate Sets" Show the Power of Chance in Comparison

Cognate Sets in Final Iteration

Cognate Sets for Semantic Leeway of 10

L1		L2		L3		L4		L5		L6		L7		L8		L9		L10		*C-C	# of L
pet	943	pet	939	0	0	0	0	0	0	bel	943	0	0	0	0	bal	936	0	0	b-l	4
pes	965	pas	961	0	0	0	0	bus	955	0	0	bas	953	0	0	0	0	0	0	b-s	4
tat	908	tut	906	0	0	0	0	0	0	0	0	dat	909	0	0	dit	906	0	0	d-t	4
tes	926	tis	922	0	0	0	0	0	0	0	0	0	0	ɟuʃ	921	daʃ	924	0	0	ɟ-ʃ	4
mes	5	mus	1	0	0	0	0	mis	5	0	0	mas	998	0	0	0	0	0	0	m-s	4
mes	3	mus	1	0	0	0	0	mis	5	0	0	mas	998	0	0	0	0	0	0	m-s	4
mat	970	0	0	mat	969	0	0	myt	963	0	0	0	0	0	0	0	0	mat	958	m-t	4
mat	970	0	0	mat	969	0	0	mat	962	0	0	0	0	0	0	0	0	mat	958	m-t	4
mit	966	0	0	mat	969	0	0	myt	963	0	0	0	0	0	0	0	0	mat	958	m-t	4
mit	966	0	0	mat	969	0	0	mat	962	0	0	0	0	0	0	0	0	mat	958	m-t	4
0	0	tik	966	tak	968	0	0	tik	972	0	0	0	0	0	0	t'iq'	966	0	0	t'-q'	4
tip	921	tap	912	0	0	0	0	0	0	təb	905	0	0	0	0	tab	910	0	0	t-b	4
tip	921	tap	912	0	0	0	0	0	0	təb	905	0	0	0	0	tab	909	0	0	t-b	4
pes	965	pas	961	0	0	0	0	0	0	vəs	955	bas	953	0	0	0	0	0	0	v-s	4
kew	994	0	0	0	0	kuw	1	kym	998	0	0	0	0	0	0	χaw	991	0	0	χ-w	4
pet	943	pet	939	0	0	0	0	bod	930	0	0	0	0	0	0	0	0	0	0	b-d	3
pit	935	pet	939	0	0	0	0	bod	930	0	0	0	0	0	0	0	0	0	0	b-d	3
pet	943	pet	939	0	0	0	0	buɟ	943	0	0	0	0	0	0	0	0	0	0	b-ɟ	3
pip	969	pop	962	0	0	0	0	0	0	bef	957	0	0	0	0	0	0	0	0	b-f	3
pokw	957	pekw	948	0	0	0	0	0	0	0	0	0	0	0	0	bik'w	947	0	0	b-k'w	3