A MODEL OF LANGUAGE EXTINCTION AND FORMATION[1]

Richard Roark

This paper reports the attempt to construct a theoretical model of the process whereby languages, during the passage of time, expand into neighboring geographical areas, replace languages formerly spoken, and ultimately split into separate languages. In addition, a comparison is made of the data yielded by the model with empirical data derived from Sydney M. Lamb's 1959 classification of North American Indian Languages.[2]

Three basic assumptions underlie the construction of this model. The most important is that there is a maximum possible area (MPA) which a language can occupy and still remain one language, though the exact size is not constant for all languages. Various forces of linguistic divergence are continually operating, and if expansion occurs beyond this MPA, contact between the members of the speech community is no longer sufficient to counteract these forces of divergence. Once the boundaries of the MPA have been exceeded, it is but a question of time before the various dialects of the original languages become mutually unintelligible. The period necessary for a language which has expanded beyond the MPA to split into daughter languages may be something like six centuries.[3]

Today the MPA of some languages, such as English, is very large, due to the efficiency of modern means of communication and transportation. This efficiency is, however, a historically recent phenomenon. During most of the past, transportation and communication have presumably been at a low and rather stable level, leading to the conclusion that the MPA must have been of moderate and roughly constant size.

The second assumption is that for an area of continental dimensions the total number of languages has remained roughly constant for large periods of time (perhaps in excess of ten or fifteen thousand years). Indeed, up to a certain point in the past it may be that the number of languages spoken in a given large area increases. The reason would be that the poorer means of communication between speech communities and within a single speech community would favor linguistic diversification.

Whenever and however language originated it presumably conferred on its speakers advantages which helped them increase in numbers and occupy larger areas. It would seem likely that once language originated in, or was introduced to, a given continental area, the speakers expanded, exceeding the MPA limits in fairly rapid succession. As they did so, divergence would have set in and produced many daughter languages. Eventually most of the continent would have been occupied, so that expansion could take place only at the expense of the language into whose territory expansion occurred. The expanded language would diversify while the other language would be displaced, or its speakers would learn the new language--or perhaps even be killed by the speakers of the expanded language.

At this point, extinction of a language would normally accompany the formation of each new language, and the total number of languages would remain roughly constant, so long as the MPA remained roughly constant. In quite modern times improved means of communication have increased the limits of the MPA enormously. A few languages have expanded at the expense of many (English at the expense of many Amerindian languages, for example) and one might predict that the next several centuries will witness a sizable decline in the number of languages spoken in the world.

The third assumption is that in any six century period there is some probability that a language will expand beyond the limits of the MPA. The actual value was initially assumed to be 50 percent that tentative expansion (to be explained in the operating rules) would occur.

In order to understand the processes of language extinction and formation it may be useful to consider one MPA at two separate points in time.

I. The language initially in the MPA may no longer be spoken.
   A. The speakers may be dead.
   B. The language may have been replaced by another language.
   C. The speakers may have moved elsewhere. (Note that such removal would not alter the total number of languages spoken.)

II. The language may still be one language.
   A. Its area may have contracted.
   B. Its area may have expanded without exceeding the MPA limit.
   C. Its area may have remained constant.
   (None of these possibilities would alter the total.)

III. The initial language may now be more than one language. Such diversification will presumably occur six centuries after the MPA limit is exceeded.

Since the total number of languages is assumed to remain constant, $I_A + I_B = III$.

In summary there seem to be two basic processes leading to language formation and two leading to language extinction.

I. New languages may be formed by
   A. Spontaneous generation. A language may appear with no prior glottogene. (By glottogene is meant a language viewed through time, rather than at an instant in time.) This process must have occurred at least once, but has never to this writer's knowledge been observed, and no provision for it has been made in the model.
   B. By diversification of languages which have expanded beyond the MPA limit.

II. Languages may become extinct
   A. Because the speakers learn another language. This process will normally be due to intrusion of a foreign language. (I_B)
   B. Because the speakers have all died. This process might or might not be due to intrusion of the speakers of a foreign language.

The model attempts to represent these processes by means of a series of squares, representing the MPA's of an entire continent. Most of these squares will contain a language; a few will be blank. Surrounding the continent is a row of border squares which limits expansion to the confines of the continent. This border may be thought of as the ocean. Whenever a language expands from one MPA into another it replaces the previous language, and then diversifies to produce one or more new languages six centuries later. Diversification also occurs when a language expands into a blank square, and in addition a language chosen at random becomes extinct, so that the total remains constant.

The model consists of 40 continental squares (8 x 5 on each side) surrounded by a row of border, or ocean squares. The squares are numbered (n) from 1-70. Initially 32 of the 40 continental squares are assigned a language; the other 8 are left blank. Each square bears a Replacement Number (RN) in the upper left hand corner and a Diversification Number (DN) in the lower right hand corner. The RN of each ocean square is always 33. The RN's of other squares, and all DN's have two parts (a,b).

Other components of the model are:

1. Expansion Probability Selector: 16 numbers, eight of which are zero. The others are -8, -7, -6, -1, 1, 6, 7, 8 (the numbers of squares contiguous to any given square). This selector provides an 8 out of 16 chance that tentative expansion will occur into a given square during a six century period.
2. Random Extinction Selector: contains the numbers of all continental squares, allowing a language chosen at random to become extinct when expansion takes place into a blank square (numbers are 9-13, 16-20, 23-27, 30-34, 37-41, 44-48, 51-55, 59-62).
3. Random Extinction File (REF): room for the numbers of languages which become extinct at random (up to a maximum of eight).
4. Don't Use Again File (DUA File): sublists 1-32. Each sublist contains a number (d,u). Initially these numbers are 1, 0; 2, 0; 3, 0 . . . 32, 0. This file ensures that no language can be produced by diversification more than once.
5. First Occurrence File (FOF): room for 32 numbers (F,O). This file keeps track of the first occurrence of a language in each six century period. (The first occurrence is arbitrarily defined as the one in the square with the lowest number.)

The model starts at the beginning of a six century period with one language per MPA. Immediately after the beginning of the period expansion takes place, resulting in some languages occupying more than one MPA. By the end of the six century period divergence results in the formation of new languages (n-1 for each n MPA's occupied by a single language).

General Expansion and Diversification Rules

1. Expansion takes place from a square containing a language to a contiguous, non-ocean square, subject to certain restrictions.

2. Expansion takes place only once into a given square per six century period.

3. If expansion attempts to take place from square A to square B and also from B to A during the same period, no expansion results.

4. Expansion into an occupied square results in the extinction of the language previously there.

5. Expansion into a blank square is followed by extinction of a language chosen at random.

After expansion, languages which occupy more than one MPA diverge for the remainder of the six century period. At the end of this period they become separate languages.

6. When languages occupy more than one MPA the first instance of the language is not altered. Each succeeding instance becomes a new language.

7. Diversification may not result in the production of a language presently or previously in existence.

Operating Procedure

Assign RN of 33 to each ocean square. Assign RN's of 1, 0; 2, 0. . . 32, 0 at random to 32 of the continental squares. Assign RN of 0, 0 to other eight continental squares. Assign DN of 0, 0 to all squares. Assign d, u of 1, 0 to sublist 1 of DUA File; 2, 0 to sublist 2; etc. Place cards containing the numbers -8, -7, -6, -1, 1, 6, 7, 8, 0, 0, 0, 0, 0, 0, 0, 0 in the Expansion Probability Selector and shuffle. Place 40 cards, each containing the number of a continental square in the Random Extinction Selector and shuffle. Place 0, 0 in each of the eight consecutive cells of the Random Extinction File. Place 0, 0 in each of the 32 consecutive cells of the First Occurrence File. The model is now ready to operate; a picture of it is shown on the following two pages. 0, 0 is indicated by a blank.

START WITH SQUARE n = 9 and ask question ONE.

1. Can expansion tentatively occur into n? To find out, choose one of the cards in the Expansion Probability Selector. This card will have a number, q. Ask: is q zero?

   If yes, there will be no expansion. Increase n by one and ask 1.
   If no, there may perhaps be tentative expansion. Ask

2. Is n ocean? (Is RN of n 33?)

   If yes, increase n by one and ask 1.
   If no, ask

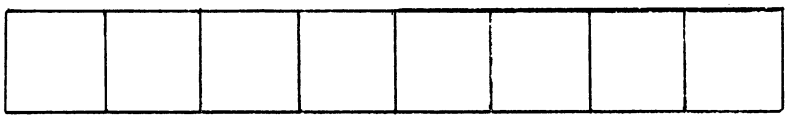
3. Has n been previously stored in the Random Extinction File? (To find out examine the first number, r, in REF and ask: is r = n?).

   If yes, increase n by one and ask 1.
   If no, increase r by one and ask 3. After r = 8 ask

33 square 1	33 2					7
	1,0	2,0	4,0		14,0	14
	15,0	29,0	21,0	9,0	18,0	21
	27,0	20,0	24,0	8,0	25,0	28
		5,0	10,0	32,0	17,0	35
		16,0	28,0	13,0	22,0	42
	6,0		7,0		26,0	49
	23,0	3,0	12,0	31,0		56
	30,0		19,0			63
						70

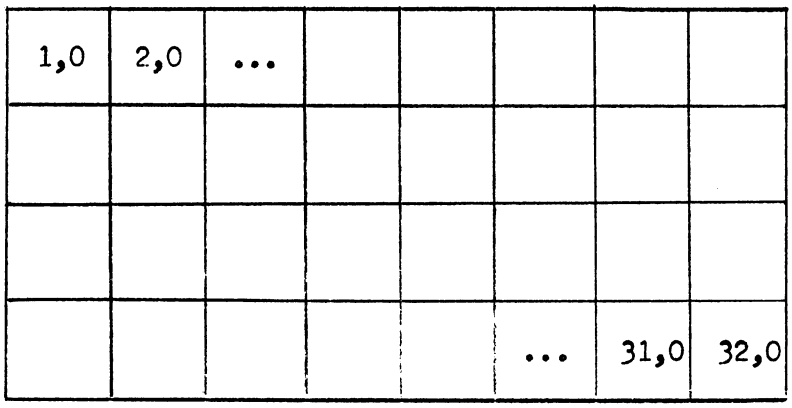
REF

CELLS 1 - 8



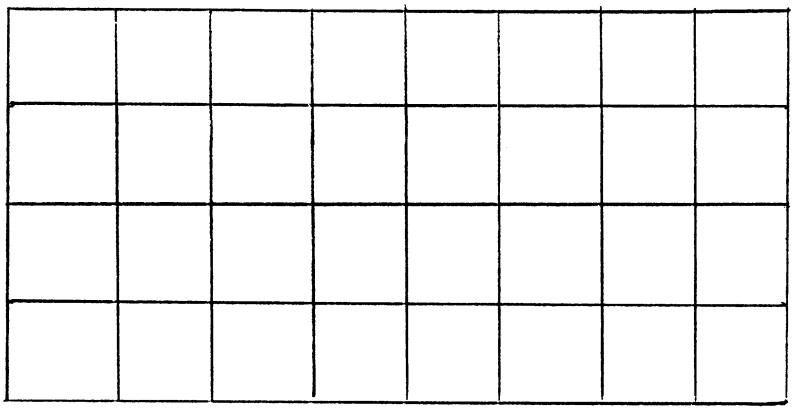
DUA FILE

sublists 1 - 32



FOF

CELLS 1 - 32



4. Is n plus q ocean?

   If yes, increase n by one and ask 1.
   If no, ask

5. Is n plus q blank? (Is RN = 0, 0?)

   If yes, increase n by one and ask 1.
   If no, there will be tentative expansion (subject to later random extinction) from n plus q to n, unless there has already been tentative expansion from n to n plus q. Ask

6. Has there been tentative expansion during this six century period from n to n plus q? (If there has, the DN of n plus q will equal the RN of n.)

   If yes, no tentative expansion takes place. 0, 0 replaces DN of n plus q, nullifying the previous tentative expansion from n to n plus q. Increase n by one and ask 1.
   If no, tentative expansion occurs from n plus q to n. DN of n replaced by RN of n plus q. Ask

7. Was n blank when expansion took place into it? (If it was, its RN will be 0, 0 at this point.)

   If no, increase n by one and ask 1.
   If yes, choose number, x, at random from Random Extinction Selector. Ask

8. Is square x blank?

   If yes, choose new x and ask 8.
   If no, store x in first zero cell in Random Extinction File. 0, 0 replaces RN and DN of x. (Language in square x becomes extinct.) Increase n by one and ask 1.

After n = 62, zero replaces each number in REF. At this point each square into which expansion has occurred will have a DN NOT EQUAL to 0, 0. Let n = 9 and ask

9. Is DN of n 0, 0?

   If yes, increase n by one and ask 9.
   If no, the new language replaces the old. DN of n replaces RN of n. 0, 0 replaces DN. Increase n by one and ask 9.

After n = 62, print picture of each continental square. All squares now have DN's of 0, 0. All occupied continental squares have RN's other than 0, 0. The same RN may exist in more than one MPA, indicating that a language has exceeded the MPA limits. The time is now just after the beginning of the six century period. This situation persists until, after six centuries of divergence, those languages which have expanded beyond the MPA have become separate languages. Let n = 9 and ask

10. Is n ocean?

    If yes, increase n by one and ask 10.
    If no, ask

11. Is n blank?

If yes, increase n by one and ask 10.
If no, it must be determined whether this is the first occurrence of the language in n. Examine the number (F, 0) in the first cell (cell f) of the First Occurrence File and ask

12. Does F, 0 = 0, 0?

If yes, this is the first occurrence of the language in n. Store RN of n in cell f of FOF. Increase n by one and ask 10.
If no, this is perhaps not the first occurrence of the language in n. Ask

13. Does RN of n = F, 0? (Note that RN is of the form a, b).

If no, increase f by one and ask 12.
If yes, this is not the first occurrence of the language in n. Replace the number (d, u) in sublist a of the DUA File by d, u plus 1. Replace RN of n by d, u plus 1. Increase n by one and ask 10.

After n = 62, 0, 0 replaces each number in the First Occurrence File. Print contents of each Continental square. The time is now just at the end of a six century period. The b part of each RN has been increased by one for each occurrence of the RN after the first. Thus the a part of each RN indicates genetic relationship and the b part indicates the past history of a group of genetically related languages. For example, an RN of 1, 4 means that this is the fourth language in a language family ultimately derived from language 1, 0.

The following pages illustrate the steps in one six century period. At the end of the period there were still 32 languages, but seven of the original 32 had become extinct, while seven had been produced by divergence from languages which expanded beyond the MPA limits. The 32 languages existing at this point go back to 25 ancestral languages which existed six centuries earlier.

Continuing the process for a total of 60 centuries yields the following data:

Time	Original Glottogenes Still Represented
0	32
6	25
12	17
18	14
24	14
30	13
36	11
42	11
48	9
54	9
60	8

Thus, viewing the situation from the present, the 32 existing languages are derived from 8 languages which existed 60 centuries ago, 13 languages which existed 30 centuries ago, etc.

The remainder of this paper compares the data yielded by the model with data obtained from Sydney Lamb's Classification of American Indian languages, previously referred to.

Dr. Lamb's classification of Amerindian languages attempts to set up genetic groups of roughly comparable time depth and roughly comparable internal diversity. Perhaps 300-400 languages were spoken in native North America at white contact; the present writer has chosen the figure 368, for computational purposes.

The classification groups these languages into 109 families, of time depth 25 to 30 centuries; 61 stocks of time depth 45 to 50 centuries, and 23 orders of time depth 65 to 75 centuries. That is, the 368 languages spoken at white contact go back to 23 glottogenes 65 to 75 centuries ago, or 60 to 70 centuries before 1500 A.D. But these 23 glottogenes did not become separate languages for perhaps six centuries more, so that the 368 languages are descended from 23 languages which existed 54 to 64 centuries ago. Letting x be centuries before 1500 A.D. and y number of languages from which the final 368 were derived:

y	x	
368	0	
109	14-19	
61	34-39	
23	54-64	(see graph one)

An equation was derived to represent this trend by choosing the values of x which led to the most regular curve.

y	x
368	0
109	19
61	34
23	64

These values yield a curve, concave to the right, when plotted on semi-logarithmic paper. A straight line on such paper indicates that there is some period, p, each iteration of which will reduce the initial value of y by one half. Such a line represents an exponential equation, and is sometimes referred to as a constant growth law. Radioactive decay is an example: there is a period, 1600 years, which will reduce any amount of radium by half. But the shape of the curve obtained from the linguistic data indicates that the period of time needed to reduce the languages spoken at white contact by half, and this number in turn by half, is itself a slowly increasing variable. Two equations were derived:

$$y = \frac{368}{(2)^{\frac{x}{\text{antilog } (.765 + .245 \log x)}}} \quad ; \quad y = \frac{368}{(2)^{\text{antilog } (.763 \log \frac{x}{10.5})}}$$

(10.5 is the number of centuries needed to reduce the original 368 languages to 184; each iteration of this period is x divided by 10.5. Thus going back

in time reduces the 368 original languages to 184, which are part of a different 368 then spoken. Going back another 10.5 centuries reduces this second 368 to 184, but only reduces the first 184 to 114, not to 92. For further discussion, refer to the mathematics appendix.)

This graph may now be compared with the results of the model. The model contained only 32 languages; that is, it was 11 and one half times smaller than Native North America. Hence all the values it produced for number of languages need to be multiplied by 11.5.

Time	Languages Produced by Model	11.5 x Languages Produced by Model	Languages Read from Graph
0	32	368	368
6	25	288	250
12	17	196	170
18	14	161	115
24	14	161	88
30	13	150	72
36	11	127	58
42	11	127	48
48	9	103	37
54	9	103	32
60	8	92	23

(See graph two)

There is similarity in shape between the theoretical and empirical curves. The overly high values produced by the model indicate that the tentative expansion probability figure of 50% is too low. Another trial was made with three zeroes in the Expansion Probability Selector, rather than eight. The following results were obtained.

Time	Languages Produced by Model	11.5 x Languages Produced by Model
0	32	368
6	22	253
12	21	241
18	16	184
24	15	172
30	12	138
36	11	127
42	10	105
48	8	92
54	7	81
60	6	69

(See graphs 3 and 4)

These figures agree more closely with the empirical data. It is to be expected that the same expansion probability figure will yield a smaller number of languages with a larger model, since the proportion of ocean squares to continental squares is less. That is, a smaller percentage of the total continental squares suffers the disadvantage of having expansion

time = 0

33 square 1	33 2					7
	1,0	2,0	4,0		14,0	14
	15,0	29,0	21,0	9,0	18,0	21
	27,0	20,0	24,0	8,0	25,0	28
		5,0	10,0	32,0	17,0	35
		16,0	28,0	13,0	22,0	42
	6,0		7,0		26,0	49
	23,0	3,0	12,0	31,0		56
	30,0		19,0			63
						70

time = 0+ expansion occurs

33 square 1	33 2					7
	~~1,0~~- (REF)	2,0 4,0	4,0	4,0	14,0	14
	15,0	29,0	~~21,0~~- (REF)	9,0	18,0	21
	27,0	20,0 27,0	24,0 21,0	8,0	25,0 8,0	28
		5,0	10,0 20,0	32,0 10,0	17,0	35
		16,0 28,0	28,0 32,0	~~15,0~~- (REF)	22,0	42
	6,0		7,0 28,0		26,0 22,0	49
	23,0 3,0	3,0	12,0	31,0 7,0	31,0	56
	30,0	23,0	19,0			63
						70

time = 0 - 6 expansion complete, slow divergence

square 1	2					7
		4,0	4,0	4,0	14,0	14
	15,0	29,0		9,0	18,0	21
	27,0	27,0	21,0	8,0	8,0	28
		5,0	20,0	10,0	17,0	35
		28,0	32,0		22,0	42
	6,0		28,0		22,0	49
	5,0	3,0	12,0	7,0	31,0	56
	30,0	23,0	19,0		11,0	63
						70

time = 6 divergence has resulted in new languages.

square 1	2					7
		4,0	4,1	4,2	14,0	14
	15,0	29,0		9,0	18,0	21
	27,0	27,1	21,0	8,0	8,1	28
		5,0	20,0	10,0	17,0	35
		28,0	32,0		22,0	42
	6,0		28,1		22,1	49
	3,0	3,1	12,0	7,0	31,0	56
	30,0	23,0	19,0		11,0	63
						70

rendered impossible from three or four contiguous squares by the presence of the ocean.

The Rules for operating the model have been written as a series of yes or no questions in order to facilitate eventual writing of a computer program which would enable the IBM 704 to run the model with any desired number of languages. Nothing more complicated than a two way program branch would need to be employed, and it is hoped that ultimately the program can be completed.

ENDNOTES

1. I am greatly indebted to Sydney Lamb for his insight and patience during the preparation of this paper, and to Dell Hymes for his critical review. Any faults remaining should be attributed to my own stubbornness.
2. Sydney M. Lamb, Some Proposals for Linguistic Taxonomy, Anthropological Linguistics 1:33-49 (1959).
3. Sydney M. Lamb, personal communication.

BIBLIOGRAPHY

Lamb, Sydney M.

1959 Some Proposals for Linguistic Taxonomy, Anthropological Linguistics 1:33-49.

MATHEMATICS APPENDIX

For computational purposes the following values of x (centuries, counting backward from 1500 A.D.) and y were chosen, as they gave the most regular possible curve on the graph:

y (languages	x (centuries
368	0
109	19
61	34
23	64 (see graph 5)

These values plotted on semilog paper yield a curve concave to the right. A straight line would represent a simple exponential equation, in which there is some period, p, which reduces the value of y by 1/2. But the shape of the curve indicates that this period is a variable which increases with time. From the graph interpolations can be made to find the time needed for successive reductions of the initial value of y by 1/2:

y	x	n
368	0	0
184	10.5	1
92	23.5	2
46	42.5	3
23	64.0	4
⋮	⋮	⋮
$\frac{368}{2^n}$	$(n)(p_{average})$	n

$p_{avg} = (p_1 + p_2 + \ldots + p_n) \div n$; $p_1 = 10.5$; $p_2 = (23.5 - 10.5) = 13$; etc.

Hence, $p_{avg} = \frac{x}{n}$ $\therefore$ $n = \frac{x}{p_{avg}}$ and $y = \frac{368}{(2)^{\frac{x}{p_{avg}}}}$

x and p_{avg} were found to be nearly in a straight line when plotted on logarithmic paper. Hence $\log p_{avg} = a + b \log x$ or $p_{avg} = \text{antilog}\ (a + b \log x)$

p_{avg}	x
10.5	10.5
12.75	23.5
14.17	42.5
16.0	64.0

(see graph 6)

$P = \log p_{avg}$; $X = \log x$

P	X	XP	X^2
1.022	1.022	1.045	1.045
1.106	1.371	1.528	1.880
1.151	1.628	1.872	2.647
1.204	1.806	2.175	3.259
4.483	5.827	6.620	8.831

n is four cases

$\Sigma(P) = na + b\Sigma(X)$; $\Sigma(XP) = a\Sigma(X) + b\Sigma(X^2)$ hence $a = .765$; $b = .245$

$p_{avg} = \text{antilog}\ (.765 + .245 \log x)$

$$y = \frac{368}{(2)^{\frac{x}{\text{antilog}\ (.765 + .245 \log x)}}}$$

This equation represents the trend of the curve moderately well, but is not trustworthy for small values of x; less than five, for example.

It will be noted that each iteration of the period 10.5 will reduce y by a successively smaller percentage.

iterations	y	$y = 368 \div 2^{\text{exponent}}$
1. (10.5)	184	2^{1}
2. (21.0)	114	$2^{1.69}$
3. (31.5)	76	$2^{2.31}$
4. (42.0)	48	$2^{2.91}$
5. (52.5)	35	$2^{3.38}$
6. (63.0)	24	$2^{3.94}$

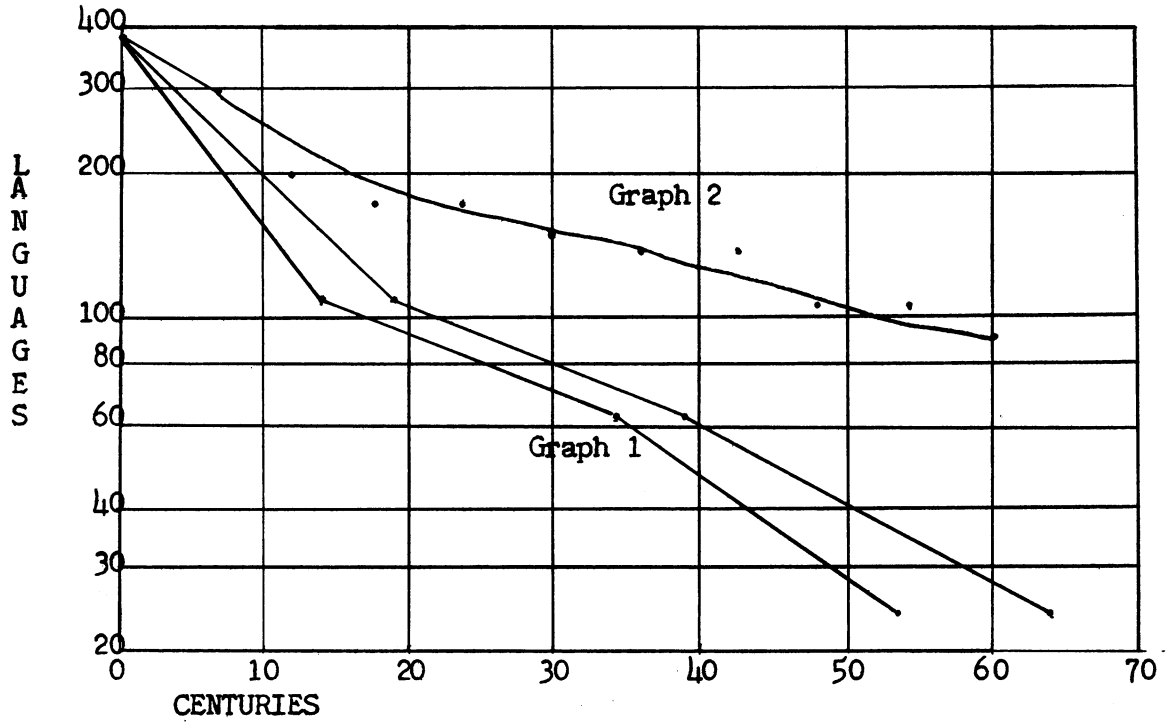
Plotting the exponent of 2 against iterations of 10.5 yields a straight line on logarithmic paper (see graph 7).
Hence log exponent = a + b log iterations. Let E be log exponent and I log iterations.

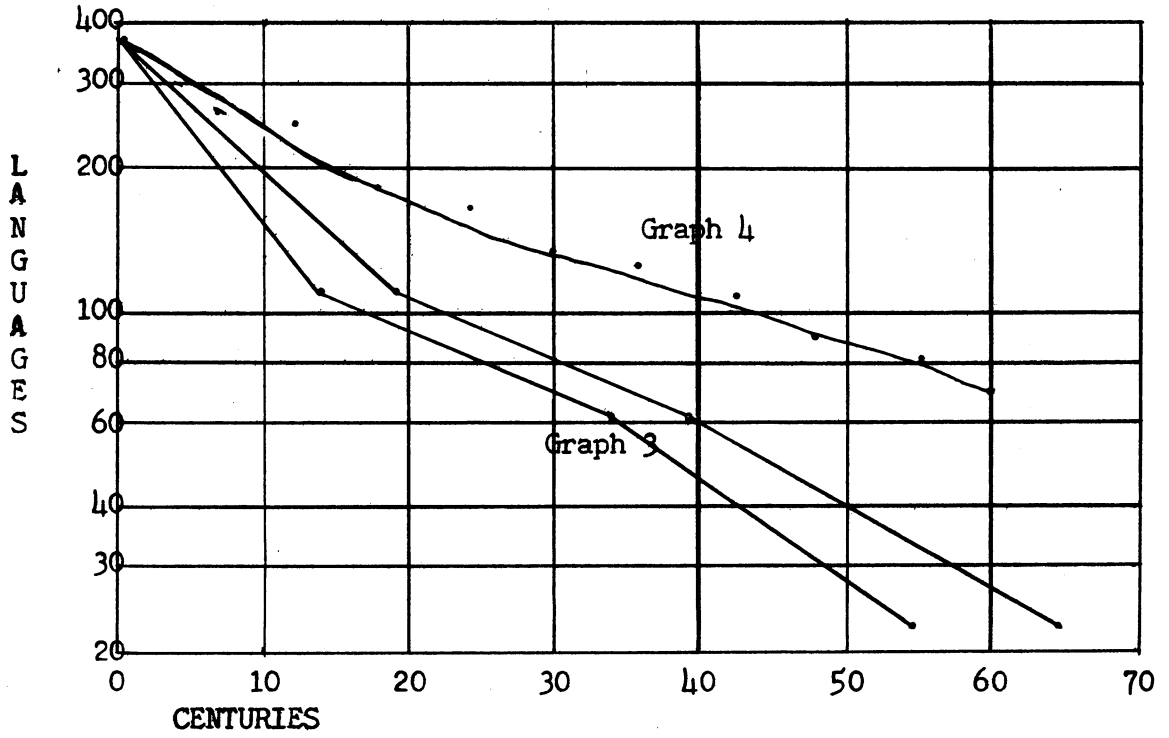
E	I	EI	I^2	
0.0	0.0	0.0	0.0	
.229	.301	.0686	.0907	
.363	.477	.1733	.227	
.464	.602	.2795	.363	n = 6 cases
.529	.699	.3700	.489	
.596	.778	.4640	.606	
2.180	2.857	1.3554	1.7757	

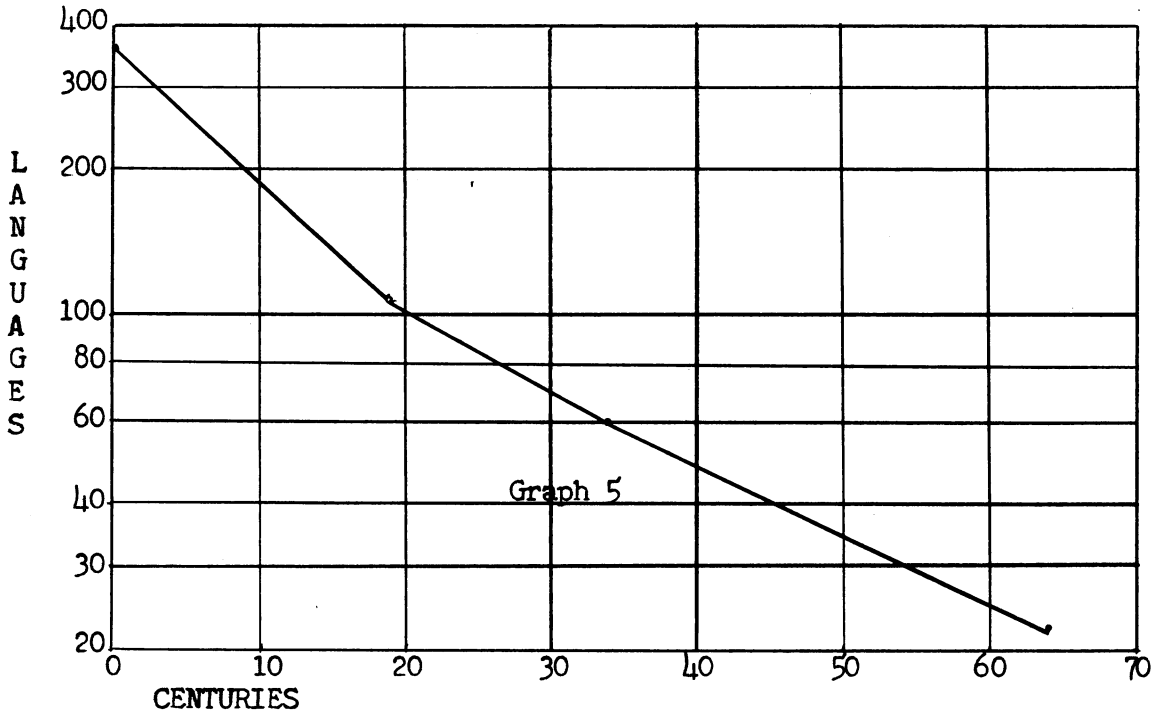
$\Sigma(E) = na + b\Sigma(I)$; $\Sigma(EI) = a\Sigma(I) + b\Sigma(I^2)$

a = 0; b = .762. Hence exponent = antilog (.762 log iterations)

$$y = \frac{368}{(2)^{\text{antilog}\,(.763\ \log \frac{x}{10.5})}}$$

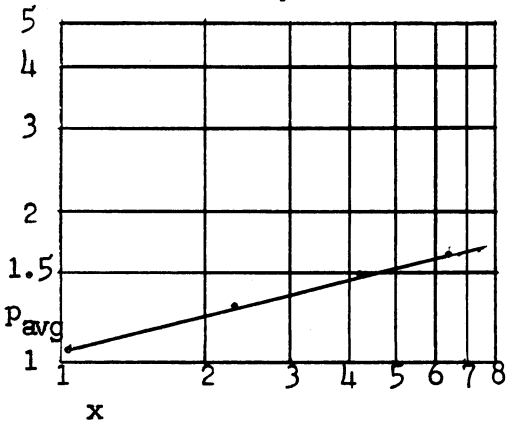

LANGUAGES
400
300
200
100
80
60
40
30
20
0
10
20
30
40
50
60
70
CENTURIES
Graph 2
Graph 1


LANGUAGES
400
300
200
100
80
60
40
30
20
0
10
20
30
40
50
60
70
CENTURIES
Graph 4
Graph 3



Graph 5

Graph 6



Graph 7