Language Polygenesis: A Probabilistic Model

Technical Report No. 435, to appear in *Anthropological Science.*

by

David A. Freedman
Department of Statistics

and

William S-Y. Wang
Project on Linguistic Analysis

University of California
Berkeley, CA 94720

Abstract. Monogenesis of language is widely accepted, but the conventional argument seems to be mistaken; a simple probabilistic model shows that polygenesis is likely. Other prehistoric inventions are discussed, as are problems in tracing linguistic lineages.

Language is a system of representations; within such a system, words can evoke complex and systematic responses.[1] Along with its social functions, language is important to humans as a mental instrument. Indeed, the invention of language—that is, the accumulation of symbols to represent emotions, objects, and acts—may be the most important event in human evolution, because so many developments follow from it. For example, Edward Sapir speculated that some embryonic form of language must have been available to early man to help him fashion tools from stone (Sapir, 1921). Sophisticated biface stone tools date to early *Homo erectus* some 1.5 million years ago, suggesting a similar age for language.

This paper considers whether the invention of language occurred at only one prehistoric site or at several sites. In other words, did language emerge by monogenesis or polygenesis? Early thinkers believed in monogenesis, against a background of divine creation. Perhaps the best known account is the biblical story of Adam giving names to plants and animals in the Garden of Eden.[2] Similar legends are found among many peoples.

Modern linguists too assume monogenesis, but on probabilistic grounds (see, for instance, Southworth and Daswani, 1974, p.314). The argument seems to be that the invention of language is an extremely unlikely event, because symbolization involves difficult abstraction and requires synchronized insight by several individuals; therefore, the probability of occurrence at more than one site must be vanishingly small. We have found no explicit quantitative treatment of this question in the literature, but the underlying logic has to be the multiplication of probabilities. If p is small at one site, then $p \times p$ for two sites is smaller still, and so on. This reasoning is false, as we show here. The fallacy lies in the focus on two *particular* sites rather than consideration of *all* pairs of sites.

We consider the period in early pre-history, perhaps 1–2 million years ago, during which language could have emerged. Our model has n sites; at each site, language emerges independently with a small probability p, integrated over the entire time period.[3] (How instantaneous probabilities vary over time is well beyond our scope, and does not affect the argument.) In this model, the number of sites at which language emerges is Binomial(n, p), which is approximately Poisson; the expected number of sites at which lan-

guage emerges is $\lambda = np$. (Technical details will be found in an Appendix.) We consider three cases: (i) λ is small; (ii) λ is moderate; (iii) λ is large.

Table 1. Poisson model. Column 1 gives the expected number of sites at which language emerges. Column 2 gives the probability of monogenesis; column 3, the probability of polygenesis. Column 4 gives the probability of language emergence, whether by monogenesis or polygenesis. (Thus, column 4 is the sum of columns 2 and 3; in the table, columns are rounded independently, so the equality is approximate.) The last column is the conditional probability of polygenesis, given that language has emerged (column 3 divided by column 4).

Expected Number of Sites (1)	Monogenesis (2)	Polygenesis (3)	Emergence (4)	Polygenesis Given Emergence (5)
0.1	0.09	0.005	0.095	0.05
0.2	0.16	0.02	0.18	0.10
0.5	0.30	0.09	0.39	0.23
1.0	0.37	0.26	0.63	0.42
2.0	0.27	0.59	0.86	0.69
5.0	0.03	0.96	0.99	0.97

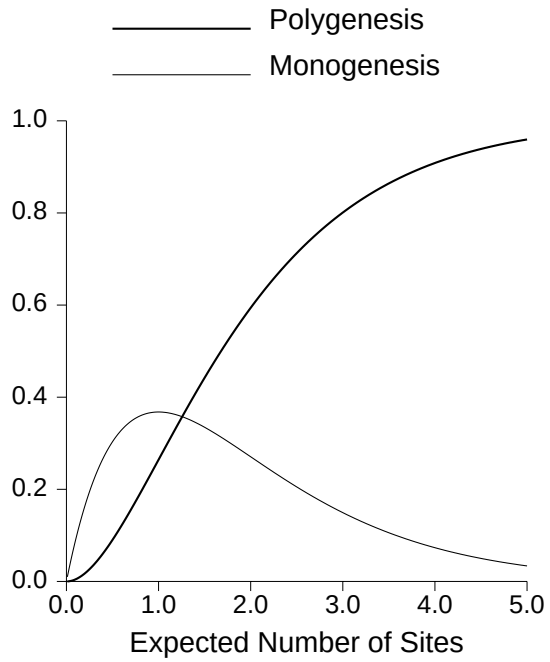
(i) If the expected number of sites λ is small, then the chance of emergence at one site is small but dominates the chance of emergence at several sites. However, we have then witnessed a rare event—the emergence of language at any site at all. In short, with these parameters, the model does not account for the fact that language has emerged. Of course, rare events do occur: statistical modeling cannot rule out monogenesis as a possibility.

(ii) If λ is moderate, so that emergence of language is not unlikely, then polygenesis is more or less as plausible as monogenesis. For example, take line 4 of Table 1. If $\lambda = 1$, the chance of monogenesis is about 37% while the chance of polygenesis is 26%, as shown by columns 2 and 3. With these parameters, the chance of language emerging—whether by monogenesis or polygenesis—is 63% (column 4). The remaining 37% probability is for the event that language does not emerge. Given that language has emerged, the probability of polygenesis is 42% (column 5).

(iii) If λ is at all large, then polygenesis is much more likely than monogenesis. For example (last line of Table 1), if $\lambda = 5$, the chance of monogenesis is only about 3%, while the chance of polygenesis is 96%.

Figure 1 summarizes the argument. If the expected number of sites is small, the chance of emergence—whether by monogenesis or polygenesis—is also small. Otherwise, polygenesis is a viable theory. In particular, monogenesis does *not* follow from the fact that p is small. Proponents of monogenesis would need to demonstrate something much stronger—that the expected number of sites is small. Moreover, such a theory has an unpalatable consequence: in effect, the emergence of language is explained as the result of a miraculous event.

Figure 1. Poisson model for language emergence. The heavier curve shows the probability of polygenesis; the lighter curve, monogenesis. Probability is a function of the expected number of sites at which language emerges (horizontal axis).



The fallacy in the conventional line of reasoning can now be seen more clearly. The chance of emergence at any particular pair of sites is indeed very small. But there are many pairs, even more triplets, and so forth. That is why polygenesis is likely. Our model is simple—indeed, artificial. Its only merit is to demonstrate the fallacy in the conventional reasoning, and show that the issue of polygenesis cannot be settled on the basis of *à priori* statistical reasoning.

While the invention of language was an extraordinary event, there do seem to be inventions that are somewhat comparable—fire, agriculture, writing. Archeological evidence suggests that these other inventions oc-

curred independently in widely scattered regions of the world. For example, pollen analysis tells us that there were at least a dozen sites where plants were independently domesticated some 10,000 years ago, ranging from New Guinea to South America (Byrne, 1987). Similarly, there is strong evidence to show that writing emerged independently at sites in Sumeria, China and Central America (Senner, 1989). From this perspective, a polygenetic scenario for language emergence does not seem so implausible.[4]

The term "language" used in this paper does not refer to the elaborate systems that we find in the world today, which are products of many millennia of development. Major transitions are needed to get from primitive symbol systems to modern languages.[5] Segmental phonology and hierarchic syntax are of particular importance (Wang, 1991, pp.105–30). However, even if we restrict attention to language systems with segmental phonology or hierarchic syntax, polygenesis is as likely as monogenesis. The whole argument remains about the same; individual probabilities may be smaller, but the number of possible sites should be larger, reflecting the growth in human populations.

Different languages in distant parts of the world have many words in common, with similar sounds and meanings. This discovery suggests that modern languages share some deep relations, and prompts the effort to trace their early lineage. However, the relations are not yet well understood and there is much controversy about lineages. For discussion of global etymologies, and a review of the controversies in tracing lineages, see Ruhlen (1995).

The probabilistic model we discuss here is unrelated to those controversies, and our argument does not depend on tracing language evolution back to first emergence. The time frames for the model and for lineage tracing are very different—1 to 2 million years versus 5,000 to 20,000 years. We believe there is little prospect of succesfully tracing lineages back to first emergence. If current thinking is correct, much of the world was colonized by waves of *Homo sapiens* coming out of Africa less than 200,000 years ago; most of the ancient symbol systems must have been replaced by the more advanced languages of the conquerors.

Studies of language contact demonstrate that any feature can be transmitted across languages. Words are borrowed by one language from another; so are sounds and grammatical constructions. Such inter-language imitations typically drive structural change and enrich the ways that languages meet new cultural needs. In consequence, no feature of a language can provide unambiguous information regarding its genetic source.

Language contacts are one result of population movements. Over a span of 200,000 years, many geophysical events have taken place that cause migrations on a massive scale, including glacial cycles, earthquakes, and volcanic eruptions. It is extremely unlikely that we can recover the lineages of either populations or languages across such migrations, thereby imposing an upper bound of perhaps 20,000 years on language histories.[6] This is an optimistic estimate; according to some authorities, "languages change at such a rate that after more than three or four thousand years of separation genetic links are no longer recognizable."[7] Our model gives no informa-

tion about time scales, but shows the plausibility of language emergence at multiple sites rather than a single site.

Notes:

1. Words may be spoken, written, or gestured; complex languages have been developed using each of these modalities. For discussion of the special relationship between brain and language, see Deacon (1992); also see Falk (1992).

2. *Genesis* 2:19. Accoding to the same source, language first diversified in the Tower of Babel, *Genesis* 11:6.

3. In China alone, well over a dozen widely-scattered sites are known for *Homo erectus* and early *Homo sapiens*, settled over a million years ago: see Wu and Olsen (1985). Additional sites are continually being discovered; for instance, hominid teeth unearthed at Longgupo date back almost two million years. See Huang et al. (1995). This archeological evidence supports the assumption that n is large.

4. There is some evidence for parallel evolution of vocalization in nonhuman primates: for a review, see Macedonia and Evans (1993). This development may be similar to parallel evolution of language. Of course, some authorities may reject the analogies we draw between emergence of language and emergence of fire, agriculture, or writing. In any case, after emergence—whether by monogenesis or polygenesis—the spread of language by diffusion seems highly probable, while many independent lineages may have become extinct.

5. There are some 5,000 languages in the world today (Ruhlen, 1991).

6. In particular, commonalities of words and sounds cannot settle the question of monogenesis versus polygenesis, because lineages cannot be traced back in time nearly far enough to matter, and because there are multiple explanations for these commonalities. This argument is made, for instance, by Dolgopolsky (1995).

7. (Dixon, 1980, p.237). Other estimates are more generous than Dixon's; so far, nothing conclusive can be said about any of these estimates; however, 5,000–20,000 years may be a reasonable range. Our span of 200,000 years for large-scale migrations could reasonably be reduced to 100,000 years. Such changes would not affect the argument very much.

Acknowledgment. We thank Luigi Luca Cavalli-Sforza, Christopher Meacham, Merritt Ruhlen, Vincent Sarich, P. Thomas Schoenemann, and Barbara Tversky for useful discussions. We also thank the editor and two referees for helpful comments.

References

Byrne, R. (1987) Climatic change and the origins of agriculture. In *Studies in the Neolithic and Urban Revolutions.* (Manzanilla, L., ed.) pp.21–34

Deacon, T. W. (1992) Brain-language coevolution. In *The Evolution of Human Languages.* (Hawkins, J. A. and Gell-Mann, M., eds.) pp.49–83. New York: Addison-Wesley.

Dixon, R. M. W. (1980) *The Languages of Australia.* Cambridge University Press.

Dolgopolsky, A. (1995) Linguistic Prehistory. *Cambridge Archeological Journal* 5: 268–271.

Falk, D. (1992) *Evolution of the Brain and Cognition in Hominids.* New York: American Museum of Natural History.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications,* vol. I, 3rd ed. New York: John Wiley & Sons.

Huang, W. et al. (1995) Early *Homo* and associated artefacts from Asia. *Nature* 378: 275–278.

Macedonia, J. M. and Evans, C. S. (1993) Variation among mammalian alarm call systems and the problem of meaning in animal signals. *Ethology* 93:177–197.

Ruhlen, M. (1991) *A Guide to the World's Languages.* Stanford University Press.

Ruhlen, M. (1995) *On the Origin of Languages.* Stanford University Press.

Sapir, E. (1921) *Language.* Harcourt and Brace.

Senner, W. M., ed. (1989) *The Origins of Writing.* University of Nebraska Press.

Southworth, F. C. and Daswani, C. J. (1974) *Foundations of Linguistics.* New York: Free Press.

Wang, W. S-Y. (1991) *Explorations in Language.* Taibei: Pyramid Press. Pp. 105–30.

Wu, R. and Olsen, J. W., eds. (1985) *Paleoanthropology and Paleolithic Archeology in the People's Republic of China.* Academic Press.

Appendix

Let $j! = 1 \times 2 \times 3 \times \ldots \times j$, the number of ways to arrange j things in order. Let

$$\binom{n}{j} = \frac{n!}{j!(n-j)!},$$

the number of combinations of n things taken j at a time. If a coin lands heads with probability p and is tossed n times, the chance of getting j heads—and therefore $n - j$ tails—is

$$\binom{n}{j} p^j (1-p)^{n-j},$$

which defines the binomial distribution, denoted Binomial(n, p) in the text. Suppose n is large but p is small; then the binomial distribution can be approximated by the Poisson distribution with mean $\lambda = np$. The Poisson distribution is often used to model the number of occurrences of a rare event; the chance of j occurrences is

$$e^{-\lambda} \frac{\lambda^j}{j!}.$$

For instance, if $\lambda = 5$, the chance of no occurrences is $e^{-5} \approx 0.0067$, the chance of one occurrence is $e^{-5}5^1/1! \approx 0.0337$, and the chance of two or more occurrences is

$$\sum_{j=2}^{\infty} e^{-\lambda} \frac{\lambda^j}{j!} \approx 0.9596.$$

See for instance (Feller, 1968). With a Binomial model, the sites must have a common probability p. With the Poisson approximation, probabilities may differ from site to site; the expected number of sites is the sum of the individual probabilities, $\lambda = \sum_j p_j$. Independence must still be assumed, and the individual p_j must be small. Table 1 was computed using the Poisson not the Binomial. The Poisson model can be elaborated to reflect site by time interactions, including appearance and disappearance of sites, with essentially the same conclusions; such complications seem unnecessary for present purposes.