When Ruhlen's 'mother tongue' theory meets the null hypothesis

Louis-Jean Boë†, Pierre Bessière‡ and Nathalie Vallée†

† ICP Université Stendhal, INPG, CNRS, Grenoble, France
‡ GRAVIR, CNRS, Grenoble France

E-mail: boe@icp.inpg.fr, Pierre.Bessiere@imag.fr, vallee@icp.inpg.fr

ABSTRACT

The demonstration of a relationship between languages can depend on finding words of similar phonological shape and roughly equivalent meaning. But it must be shown that the similarities observed could not have arisen by chance. That is to say, the null hypothesis can be rejected. We demonstrate, by a simple application of probability theory, that the world roots proposed for a Proto-Sapiens language by Merritt Ruhlen in *The origin of Languages* are the result of random chance. The null hypothesis can not be rejected. The author used too few roots, too many equivalent meanings, too many languages per family, and too many phonological equivalences for a too small number of different phonological shapes. Our calculating the factor of chance in multilateral language comparisons is a general procedure than can be used to test the limits of the methodology of Greenbergian mass comparisons.

1. INTRODUCTION

At present, there is a considerable literature showing in detail that criteria for determining a match both semantically and phonologically are almost entirely lacking, and there are numerous publications arguing against multilateral comparisons [1-7]. These numerous criticisms notwithstanding, Ruhlen's theory of the mother tongue based on global etymologies [8-9] is systematically presented as a fact in numerous popular scientific works or magazines (*Scientific American*, *Popular Science* or in France *La Recherche*, *Science et Avenir*) [10-12]. *The origin of Languages* is even recommended by MIT as a reference book in language and linguistics [13]. Gell-Mann, who received the Nobel prize in physics for his work on the theory of elementary particles, wrote the preface of the collective book *The Evolution of Human Language* [14]. He argued that is possible to construct a family tree for all the world's languages by analyzing similarities between them. He stated that any argumentation against this possibility is "so silly on the face of it that you wonder how adult human beings can adopt it." Some people see in the demonstration of the existence of a Proto-Sapiens mother tongue a confirmation of facts presented in the Bible: "Note how the main language branches Semitic, Turkic and Indo-European meet at just the point where the Tower of Babel may be. Not in Babylon but in the area East of the Black Sea" [15]. For almost five years, this phenomenon of propagation that Dan Sperber calls *Contagion of ideas* [16], has seemed to us important to analyze. We came up with a project to try to understand why some interdisciplinary fields, such as linguistics, are attracted by theories that are ill founded. We intend to investigate why there is continued propagation of such theories while the proof of their validity has not been established or even that they have been falsified. Our project is called *Representation and diffusion of scientific ideas in speech and language sciences* [17-18].

For centuries, theories addressing the origin of man and the origin of languages have been closely linked. In the XVIIth century the old interest in the origin of language and the identity of mankind was rekindled and became a dominant subject of discussion, but in the orthodox view, the Bible was still the main source of information about the earliest history of the earth and of mankind. It was believed that the earth, mankind and human language with it, were no older than about six thousand years. By the turn of the nineteenth century, with the simultaneous developments of comparative philology and anthropology, the question of the origin of man and language has been revisited with new perspectives. In Germany and France relationship between linguistics and anthropology was very close. It could be noted that a circular argumentation was already appearing. On the one hand linguists used anthropological hypothesis to validate their assumption on monogenesis for language. On the other hand anthropologists referred to the hypothesis of a mother tongue to corroborate the assumption of monogenesis of mankind. Nowadays, "the New Synthesis" which associates Cavalli-Sforza, a geneticist [19], Renfrew, an archeologist [20], and Ruhlen, a linguist [8-9] uses the same circular argumentation to reconstruct human evolution. In spite of the numerous criticisms leveled against the methodologies adopted by Cavalli-Sforza, Renfrew and Ruhlen, the CNRS (*French National Center for Scientific Research*, equivalent to NSF in the USA) launched a national project entitled *Origin of Man Language and Languages*. It considered that the frames elaborated at the end of the eighteen century and at the beginning of the nineteen century by the Indo-Europeanists "exploded overwhelmingly" with the coming of the "New Synthesis" works [21-22]. Being part of this nation wide project we intend to test Ruhlen's hypothesis statistically.

2. "MEGALOCOMPARISONS"

For about fifteen years now, Ruhlen's works in genetic typology of languages, based on multilateral comparisons of sound shapes and meaning similarity for all languages of the world, have tried to validate the existence of global

 ISBN 1-876346-48-5

roots. Recent advances in biological taxonomy serve to confirm this author's classifications of macro-families, and by implication, monogenesis of all languages. Ruhlen's thesis states that all the spoken languages around the world should be coming from a universal language. With *The origin of Language. Tracing the evolution of the mother tongue* [8], he gave further data supporting his thesis. According to him, his theory is backed up by a methodology which enables him to look for and find phonological and semantic equivalencies between words of different languages. In the end, these equivalencies enabled him to make comparisons from a set of 32 families. He finally proposed 27 global etymologies and, for each of these mother tongue roots, the most general meaning and the phonological shape (table 1).

1. mother older female AJA	**2.** knee to bend BU(N)KA	**3.** ashes dust BUR	**4.** nose to smell čUN(G)A
5. hold (in the hand) KAMA	**6.** arm KANO	**7.** bone KATI	**8.** hole K'OLO
9. dog KUAN	**10.** who? KU(N)	**11.** woman KUNA	**12.** child MAKO
13. to suck nurse, breast; MALIQ'A	**14.** to stay (in a place) MANA	**15.** man MANO	**16.** to think (about) MENA
17. what? MI(N)	**18.** two PAL	**19.** to fly PAR	**20.** arm POKO
21. vulva PUTI	**22.** leg foot TEKU	**23.** finger one TIK	**24.** earth TIKA
25. leg foot TSAKU	**26.** hair TSUMA	**27.** water ʔAQ'WA	

Table 1. World roots with their most general meanings, and phonological shapes.

3. "LA FURIA DELL'ETIMOLOGIA"

We would expect Ruhlen to select world roots without any semantic overlap. But his quest for equivalences led him to extend the meaning of each root. Eco coined the expression "la furia dell'etimologia" to dub this frantic etymological hunting [23]. As a result the number of etymological connections among the roots increased. Thus, the general meaning of root number 23 **'finger, one'** is associated with equivalent etymologies such as: 'finger nail', 'first', 'five', 'foot', 'guy', 'hand', 'index finger', 'merely', 'only', 'palm (hand)', 'paw', 'ten', 'to point', 'to say', 'to show', 'thing', 'toe'. The root numbers 10 **KU(N) 'who?'** and 17 **MI(N) 'what?'** share eleven meanings: 'do what', 'how many', 'what', 'what kind', 'what sort', 'when', 'where', 'who', 'who(ever)', 'why'.

It is also surprising to note that the list contains two identical roots (numbers 22 and 25), with the same general meaning **'leg, foot'**, two different phonological shapes **TEKU** and **TSAKU,** and sharing five meanings: 'ankle', 'hip', 'hoof', 'thigh', 'upper leg'.

Figure 1 presents semantic connections between the 27 roots. Each link indicates that the two roots have at least one meaning in common. After counting, we found out that 118 meanings are shared by two or more roots, and on average a root shares one meaning with three other roots. If we put together the roots which share one or more meanings we obtain four semantic groups: (1) roots 1, 9, 11, 12, 15; (2) root 14; (3) root 18; (4) roots 2-8, 10, 13, 16, 17, 19-27;. The first group refers to people: mother, woman (and bitch), man, child. The second and third correspond to an action (to stay) and a number (two), and the last is a hotchpotch due to the "etymological fit of rage". Increasing the number of semantic equivalences leads to an increase in chances to "discover" world roots.

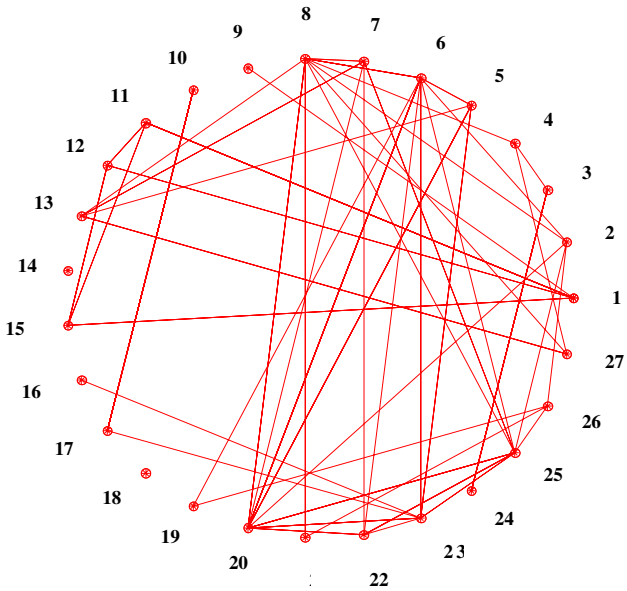

Figure 1. Connections between roots sharing at least one meaning (an average of three).

4. "COMMON SENSE" EQUIVALENCES

Ruhlen does not use complex rules to authenticate similarities between phonological shapes. Concerning consonant similarities he wrote : "You don't need a Ph.D. in linguistics (...), just common sense" [8, p. 18]. In fact, he adopted well known similarities proposed by Indo-Europeanist of the nineteenth century without taking chronology into account So he proposed three main classes of similarities [8, p. 40)]:

(1) *p* p^h *f b ɓ ß* (2) *t θ d ɗ ð* (3) *k* k^h *χ g ɣ*.

In fact he added *n, l, r* to class 2. For him vowels "*i, e, ɛ* are similar to each other, as are *u, o, ɔ*" [8, p. 39]. For the root TIK, Ruhlen considers the following phonological shapes as equivalent: *atgu, ʃiʃ, deʕ, digitu, dlianŋ, dɔkku, itygin, iʃaki, motook, otejji, řak, sik, taihwo, təgu, tku,* $ts^{\cdot}iʕ^{w}$*, sakwe, zɛkatikkuagpaa*. The vowel I of this root is judged to be an equivalent of nine other vowel phonemes as indicated in Table 2. In fact, Ruhlen considers all vowels to be similar. Increasing the number of phonological equivalences also increases the chances to "discover" world roots. Following Ruhlen's procedure, is it possible to suggest the existence of a pre-proto-sapiens

language spoken more than 50,000 years ago? This language could have consisted of 4 pre-roots, with (V)CV(C) as phonological shapes, 4 consonants (P M T K), and only one vowel V, given that all vowel qualities were equivalent.

i	e	ɛ	a	y
38%	20%	1%	15%	1%
ɨ	ə	u	o	ɔ
1%	1%	8%	13%	1%

Table 2. Vowel phonemes and percentages of occurrences. corresponding to phoneme I of the root TIK.

5. TESTING THE NULL HYPOTHESIS

Any demonstration of a relationship between languages depends largely on finding words of similar phonological shape and roughly equivalent meaning in the languages. However it must be shown that the similarities observed could not have arisen by chance. Unfortunately Ruhlen did not take this precaution. It is therefore necessary to determine if the similarities observed by Ruhlen provide a reason to reject the null hypothesis, that is, the hypothesis that they are merely a product of chance factors. "As linguists like Larry Trask, Don Ringe and Lyle Campbell, to name but a few, loudly insist, no good answer has yet been given to the charge that the correspondences noted by the long-range reconstructionists are not above the chance level. In other words, no effort has been put into rejecting the null hypothesis" [24].

A database corresponding to the global etymologies has been implemented at ICP, and the phonological segments were normalized [25]. We were thus able to assess the following parameters.
R the number of roots: 27
M the mean number of meanings per root: 24
F the total number of families: 32
N the total number of languages, proto-languages included: 1317
L the mean number of languages per family (N/F): 41
P the total number of different phonological shapes: 2739.
Based on these parameter values, we constructed "random languages" associating to each word of the language a phonological shape randomly selected (uniform distribution) among the possible P phonological shapes. This generation is applied N times to build N random languages. Then we generated F families attributing randomly each of the N languages among the F families (drawing with a uniform distribution).
We aimed at calculating, as a function of the number of the meanings of each root, the probability that each of the 27 roots appears with the same phonological shape in at least one language from each of the 32 families [26]. In fact, Ruhlen is satisfied with a less severe criterion: "the 27 world roots (...) are represented in at least six of these families, but on the average a root is represented in 12 families, and the most spread one , KU(N), 'who?', is represented in 23 or 24 families" [27, p. 234].
We propose the following procedure:

1. We associate M meanings to each of these R world roots. Each meaning corresponds to one phonological shape in each language

2. The probability P1 that a given phonological shape among the P possible has the same phonological shape in no other language L_2 of another family F_2 is calculated by:
$$P1 = (1–1/P)^{ML}$$
3. The probability P2 that no phonological shape associated to a given root R_1 of a family F_1 (there exist ML phonological shapes associated to a given root in a given family) has not the same phonological shape in any other language L_2 of another family F_2 is calculated by:
$$P2 = P1^{ML}$$
4. The probability P3 that at least one phonological shape associated with a root R_1 of a family F_1 has the same phonological shape in at least one language L_2 of another family F_2 is calculated by:
$$P3 = 1–P2$$
5. The probability P4 that a root appears with the same phonological shape in at least one language in each family is calculated by:
$$P4 = P3^{F}$$
6. Finally, the probability P5 that each of the roots R appears with the same phonological shape, in at least one language of each family is calculated by:
$$P5 = P4^{R}$$

6. RESULTS

Figure 2 displays the probability P5, as a function of the number M of meanings per root. This probability reaches 1 for five or more meanings by roots; but Ruhlen uses on average 24 meanings per root. For his derivation to be statistically founded, with the 2739 phonological shape of his database, he should have had to consider, at most, only 2 meanings per root. With 4 meanings, it appears that the probability to obtain correspondences is already above 90 %.

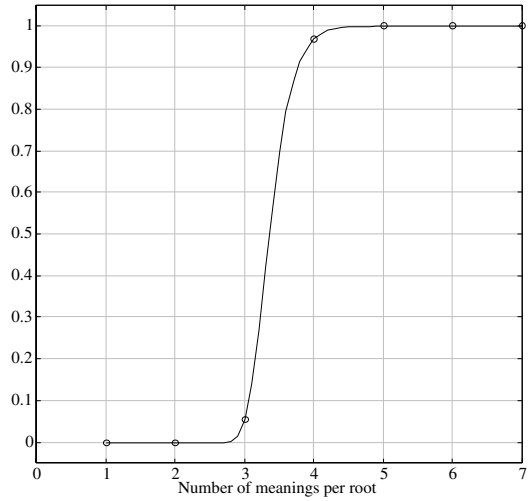

Figure 2. Probability that each root appears by chance, with the same phonological shape, in at least one language from each of the 32 families, as a function of the number of meanings per root.

ISBN 1-876346-48-5

7. CONCLUSION

The data and procedure employed by Ruhlen are far from convincing. Through a random selection of his data we have been able to reach the same conclusion as his. The null hypothesis, which stipulates that similarities among words of the world languages arise by chance, can not be rejected since its probability is equal to one! Our statistical analysis shows that with 27 roots, 24 meanings by root, 32 families and 41 languages by family, Ruhlen had 100% chance of finding world roots. With only 2739 different phonological shapes, Ruhlen should not have used more than 3 meanings per root. He used too few roots, too many equivalent meanings, too many languages per family, and too many phonological equivalences for a too small number of different phonological shapes.

One can attempt to explain why the null hypothesis can not be rejected. All speakers on the planet use the same vocal tract to differentiate phonological shapes of lexical items. There exists a common set of vowels and consonants which are used across all the world languages [28-30] and elucidated by ontogenesis [31-32]. By extending the semantic field of words it is possible to find equivalent phonological shapes across world languages. Our method of calculating the factor of chance in multilateral language comparisons can be used to test the limits of the methodology of Greenbergian mass comparisons.

Our demonstration is not a proof that the hypothesis of monogenesis for language can be rejected, but uniquely that the procedure adopted by Ruhlen to prove it has no statistical validity. It seems that the method used to reconstruct proto-languages is one of the weakest points of this "New Synthesis" which is bringing together genetic, archeological and linguistic data in the reconstruction of human evolution.

Acknowledgements

This research was conceived and started in the *Representation and Diffusion of Scientific Ideas in Speech and Language Sciences* project (managed by Louis-Jean Boë et Christian Abry, ICP, since 1999) and funded by the *Maison des Sciences de l'Homme – Alpes*. It is an ongoing research in the *Congruence* project (managed by Pierre Darlu, INSERM, since 2001), being part of the *Origine de l'Homme du Langage et des Langues* project (managed by Jean-Marie Hombert) funded by the CNRS. Thanks a lot to Christian Abry, Pierre Badin, Michel Contini, Manu Mazer, Boyd Michailovsky, Diverson Mzemba, and Anne Vilain for their assistance.

REFERENCES

[1] L. Campbell, "Review of Joseph Greenberg, Language in the Americas," *Language*, 64, 591–615, 1988.

[2] J. Guy, "Merritt Ruhlen: On the Origin of Languages, book review," *Anthropos: revue internationale d'ethnologie et de linguistique*, 90, 638–639, 1995.

[3] J. Matisoff, James "On megalocomparison," *Language*, 66, 106-120, 1990.

[4] A. McMahon, and R., McMahon "Linguistics, genetics and archaeology: internal and external evidence in the Amerind controversy," *Transactions of the Philological Society*, 93, 125-225, 1995.

[5] D.A. Ringe, "On calculating the factor of chance in language comparison," *Transactions of the American Philosophical Society*, 82, 1, 1–110, 1992.

[6] D.A. Ringe, "The mathematics of 'Amerind'." *Diachronica*, 13, 135–54, 1996.

[7] R.L. Trask, *Historical Linguistics*, London: Arnold, 1996.

[8] M. Ruhlen, *The origin of language. Tracing the evolution of the mother tongue*, New York: John Wiley & Sons, Inc, 1994.

[9] M. Ruhlen, *On the origin of languages. Studies in linguistic taxonomy*, Stanford: Stanford University Press, 1994.

[10] http://www.exploratorium.edu/exploring/language/

[11] http://www.popular-science.net/origins/indoeuro2.htm1992.

[12] http://www.sciencesetavenir.com/hs_125/originelle.html

[13] http://www.mit.edu/~ejhanna/language/langbook.html

[14] .A. Hawkins and M. Gell-Mann, The evolution of human languages, Addison Wesley Longman.

[15] http://www.biblemysteries.com/library/babelimage.htm

[16] D. Sperber, *La contagion des idées*, Paris: Odile Jacob, 1996.

[17] L.J. Boë, C. Abry, *La représentation et la diffusion des idées scientifiques dans les sciences de la parole et du langage. Erreurs et leurres comme révélateurs*, Projet Maison des Sciences de l'Homme – Alpes, Grenoble, 1999.

[18] L.J. Boë "Perception et diffusion des théories scientifiques dans les sciences de la parole et du langage," In *Percevoir : Monde et langage*., D. Keller, J.P. Durafour, J.E.P. Bonnot, R. Sock Ed., Sprimont, Belgique: Mardaga, 1999.

[19] L.L. Cavalli-Sforza, A. Piazza, P. Menozzi, J. Mountain, "Reconstruction of Human Evolution: Bringing together Genetic, Archeological and Linguistic Data," *Proc. Nat. Acad. Sciences*, 85, 6002–6006, 1988.

[20] C. Renfrew, *Archaeology and language : The puzzle of indo-european origins*, London: Jonathan Cape, 1987.

[21] http://www.cnrs.fr/SHS/Pdepart/polsc/ohll.htm

[22] J.M. Hombert, "Origine de l'Homme, du Langage et des Langues," *Rapports scientifiques de fin de deuxième année 2001-2002*. Paris: CNRS, 2002.

[23] U. Eco, *La ricerca della lingua perfetta nella cultura europa*, Roma: Laterza.

[24] J.R. Hurford, Review of Michael C. Corballis, From Hand to Mouth: the origins of language, Princeton: Princeton University Press, 2002. *Journal of Linguistics*, 39, 1, to appear. http://www.ling.ed.ac.uk/~jim/corballisrevu.html

[25] L. Métoz, N. Vallée, I. Rousset, L.J. Boë, P. Bessière, "L'hypothèse des racines universelles de Merritt Ruhlen. Évaluation statistique et analyse méthodologique," *Actes des 3e Journées d'Études Linguistiques, "Les universaux linguistiques,"* Nantes, to be published.

[26] P. Bessière, L.J. Boë, L. Métoz, I. Rousset, N. Vallée "Des formes phonétiques aux proto-formes de la langue originelle. La théorie de Meritt Ruhlen à l'épreuve des probabilités," In *Origine de l'Homme, du Langage et des Langues*, J.M. Hombert Ed. , 31-32, 2002.

[27] M. Ruhlen, *L'origine des langues*, Paris: Belin, 1997.

[28] N. Vallée, *Systèmes vocaliques : de la typologie aux prédictions*, Thèse de Doctorat de l'Université Stendhal, Spécialité, 1994.

[29] J.L. Schwartz J.L., L.J. Boë, N. Vallée, C. Abry, "Major Trends in Vowel System Inventories," *J. of Phonetics*, 25, 233-253, 1997.

[30] L.J. Boë, "Tendencies in phonological structures: the influence of substance on form," *Bulletin de la Communication Parlée*, 5, 35-55.

[31] P.F. MacNeilage, P.F., B.L. Davis, "Origin of the Internal Structure of Words," *Science*, 288, 527-531, 2000.

[32] P.F. MacNeilage, B.L. Davis, "Motor mechanisms in speech ontogeny: phylogenetic, neurobiological and linguistic implications," *Current Opinion in Neurobiology,* 11, 696-700, 2001.