Des formes phonétiques aux proto-formes de la langue originelle
Analyse méthodologique et évaluation des limites

Laurent Métoz, Nathalie Vallée, Isabelle Rousset, Louis-Jean Boë, Pierre Bessière°

Institut de la Communication Parlée
BP 25 –38040 Grenoble cedex 9, France
Tél.: ++33 (0)476 82 41 19 - Fax: ++33 (0)476 82 43 35
° Leibnitz-IMAG/INPG & Gravir-INRIA Rhône-Alpes
655 av. Europe – 38330 Montbonnot Saint Martin, France
Mél : laurentmetoz@yahoo.fr , vallee,rousset,boe@icp.inpg.fr, bessiere@imag.fr

ABSTRACT

The aim of our study is, basically, based on the study leaded by Merritt Ruhlen on world-languages-classification (classification presented in Ruh94). Proving that the methodology used by Merritt Ruhlen as well as the plenty amount of data used for the alike Greenberg's comparative method might have been a factor of constraint wouldn't be possible without the constitution of a probabilistic estimation. The application of this on Ruhlen's data allows us to highlight the fact that the demonstration leaded by Ruhlen among this book is probabilistically invalidated. We show that a drawing lots of the phonetic forms of the lexico-semantic referents gives the same results than him.

1. INTRODUCTION

Depuis une quinzaine d'années, les travaux de Merritt Ruhlen en matière de typologie génétique des langues, basés sur la recherche directe de ressemblances de formes sonores et de sens dans des items lexicaux de différentes langues, tentent de valider l'hypothèse selon laquelle une seule et même proto-langue universelle serait à l'origine de toutes les langues parlées à la surface du globe ; le mythe de la Tour de Babel est ainsi réactualisé. Dans son ouvrage *The origin of Language. Tracing the evolution of the mothertongue*, [Ruh94], Ruhlen apporte de nouveaux éléments en faveur de cette thèse. Pour l'étayer, il adopte une méthodologie qui lui permet de rechercher et de trouver des équivalences phonétiques et sémantiques entre items de différentes langues en établissant un travail de comparaison à partir d'un choix de 34 familles de langues, 1 325 langues et proto-langues (une moyenne de 49 langues par familles). Il parvient ainsi à mettre en évidence 27 racines universelles associées, en moyenne, à 24 référents secondaires qui correspondent à différents glissements sémantiques du référent principal et constituent une partie du champ lexico-sémantique de celui-ci : exemple *KANO* (référent principal pour le signifié *bras*) possède 12 référents secondaires (*aile, aisselle, avant-bras, branche, bras, coude, doigt, épaule, main, manche, patte de devant, pied*). Cette méthode d'analyse lui permet de retrouver des formes phonétiques très proches dans des langues de différentes appartenances génétiques, ce qu'il considère comme preuve d'une affiliation entre celles-ci. Mais plusieurs biais méthodologiques peuvent être mis en évidence et nous montrons à travers cette étude que le hasard donne autant de chance qu'à Merritt Ruhlen de retrouver des racines universelles à partir de son corpus. La méthodologie de Ruhlen s'appuie sur la méthode multilatérale de [Gre57] pour l'apparentement des langues. Cette méthode recherche dans un très large échantillon de langues un ensemble d'items suffisamment ressemblants. Un des postulats de Greenberg est que ces items doivent être semblables à la fois en terme de sons et de sens. Lorsqu'il s'agit de comparer entre elles des familles de langues, [Ruh94] propose d'accepter l'idée de changements sémantiques simples, tout comme il existe des changements phonétiques simples. Or, intuitivement, nous percevons les conséquences d'un tel choix méthodologique. En effet, élargir le champ lexico-sémantique d'un item va de fait augmenter la probabilité de retrouver dans plusieurs familles de langues un item phonétique commun. Ainsi, accepter de nouveaux items lexicaux, c'est aussi accepter leurs formes phonétiques. À ce cela s'ajoute une multitude d'équivalences phonétiques proposées par [Ruh94]. Par exemple, il donne comme phonétiquement équivalents :

$[p] \cong [p^h] \cong [f] \cong [b] \cong [\beta] \cong [\mathfrak{b}]$

$[t] \cong [t^h] \cong [\theta] \cong [d] \cong [\delta] \cong [\mathfrak{d}]$

Ces équivalences ne sont que deux exemples parmi beaucoup d'autres. Par conséquent, il devient plus aisé de retrouver, dans une masse toujours plus importante de données, des formes phonétiques identiques entre les familles de langues. Pour chaque racine étudiée, [Ruh97] propose, en moyenne, 24 changements sur la base de glissements sémantiques (c'est-à-dire 24 référents secondaires).

2. DONNÉES ET CHOIX MÉTHODOLOGIQUES

Une étape préalable à notre étude a consisté à implémenter les données linguistiques présentées dans la partie annexe de l'ouvrage *L'origine des langues*,

[Ruh94] trad. Franç. de *The Origin of Language. Tracing the evolution of the mothertongue*, à partir desquelles Ruhlen tire sa conclusion sur l'existence de racines communes aux langues du monde. Cette tâche a nécessité une vaste tâche d'homogénéisation phonétique des données. En effet, pour les 2 850 formes phonétiques correspondant à tous les référents secondaires des 27 racines universelles, l'Alphabet Phonétique International n'a pas été systématiquement utilisé par Ruhlen : certains items apparaissent sous forme phonologique voire orthographique. C'est donc à partir de données de la littérature que cette harmonisation a pu être effectuée.

Nous avons été amenés, lors de l'élaboration de la base de données, à opter pour plusieurs choix méthodologiques. Nous avons conservé les items phonétiques qui apparaissaient plusieurs fois pour un même référent principal même si le référent secondaire est identique car ils apparaissent dans des langues différentes. Par exemple la forme [mana] du référent principal *MANA* (signifié *rester sur place*) se retrouve pour cette même racine dans des langues des familles indo-européenne, afro-asiatique et indo-pacifique pour des référents secondaires différents. Par ailleurs cette forme phonétique est donnée dans des langues des familles Niger-Congo, Amérinde ayant pour référent principal *MANO* (*homme*) et dans la famille ouralienne pour le référent principal *MENA* (*penser à*).

Nous avons également décidé de traiter sans distinction les formes phonétiques de langues et celles de proto-langues (exemple *[man] pour la forme phonétique de *MANA* en proto-afro-asiatique). Sont aussi considérées les alternances des formes phonétiques répertoriées par [Ruh97] (exemple en proto-ouralien l'alternance des formes phonétiques suivantes pour *MANA* *[man]v~*[mon]v où v désigne une qualité vocalique), également les formes non autonomes comme [mann]- dans la famille caucasienne pour *MANA* ou [ma]- dans la famille amérinde. Ainsi 2 850 formes phonétiques différentes ont été dénombrées automatiquement après regroupement des formes identiques.

3. ÉVALUATION DES LIMITES

Après l'implémentation de la totalité des données de [Ruh97] et évaluation du nombre de catégories phonétiques et lexico-sémantiques des référents secondaires, nous avons tenté une estimation probabiliste de l'hypothèse des racines universelles. Pour ce faire, nous avons évalué la probabilité d'obtenir des formes phonétiques identiques dans plusieurs familles de langues par tirage au hasard des formes phonétiques répertoriées par l'auteur [Ruh97] .

Pour notre démonstration nous considérons les variables et paramètres suivants :

R : nombre de référents principaux correspondant à l'ensemble des racines universelles proposées par [Ruh97], par exemple : AJA pour le champ lexico-sémantique *mère parent féminin plus âgé*

r : nombre moyen de référents secondaires ; c'est-à-dire par exemple : parent féminin plus âgé, grand-mère, tante, femme, épouse, belle-mère, mère du père, vieille femme, tante paternelle, etc. référents secondaires pour AJA.

F : nombre total de familles de langues pour l'ensemble des racines universelles ; par exemple : khoisan, Niger-Congo, toungouze, dravidien, etc.

L : nombre moyen de langues par famille (proto-langues comprises). Sont pris en compte, par exemple, temne, bulom, yoruba, proto-bantou pour la famille Niger-Congo et le référent principal AJA.

P_{obs} : nombre d'items ou formes phonétiques observés pour l'ensemble des référents principaux (pour AJA, par exemple, on rencontre : [aja], [aija], [aijako], -[ja], [ja], [aj], [jaja], *[jaʔ], etc.)

P_{max} : nombre d'items phonétiques maximum , soit R*r*L

L'exploitation des données de [RUH97] nous livre les valeurs suivantes pour les variables et les paramètres :

$R = 27$

$r = 24$

$F = 34$

$L = 39$

(1 325 langues pour l'ensemble des 34 familles)

$P_{obs} = 2\ 850$

$P_{max} = R*r*L = 25\ 272$

La probabilité P qu'une forme phonétique correspondant à un référent secondaire j d'un référent principal donné appartienne à au moins une langue de chaque famille linguistique est donnée par

$$(1 - (1 - Lr/p_{obs})^{Lr})^{F}$$

avec $(1 - Lr/p_{obs})^{Lr}$ la probabilité pour un référent principal donné pour qu'aucune forme phonétique du référent secondaire i de la famille linguistique j n'appartienne aux formes phonétiques du référent secondaire i de la famille linguistique k.

Ainsi la probabilité pour l'ensemble des référents principaux sera obtenue par P^R.

4. RÉSULTATS

Les figures 1 & 2 montrent respectivement les valeurs de P et P^R en fonction du nombre r de référents secondaires (nombre moyen par référent principal). Pour P_{obs} = 2 850, dans la figure 1, on constate que P = 1 à partir de quatre référents secondaires et dans la figure 2 que $P^R = 1$ lorsque r = 5. Or [Ruh97] utilise

en moyenne 24 référents secondaires par référent principal ou racine universelle. Si on utilise cette fois P_{max} dans le calcul de P, on remarque alors que P = 1 lorsque r = 12 dans la figure 1 et r = 15 dans la figure 2.

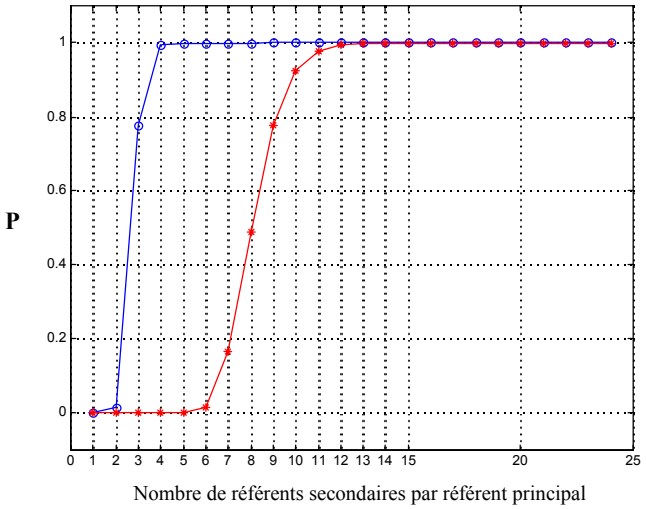

Figure 1. Valeurs de P en fonction du nombre moyen de référents secondaires par référent principal pour 39 langues en moyenne, 34 familles et pour un nombre d'items phonétiques observés (P_{obs}) égal à 2 850 (courbe o), et pour un nombre d'items phonétiques maximal (P_{max})égal à 25 272 (courbe ∗).

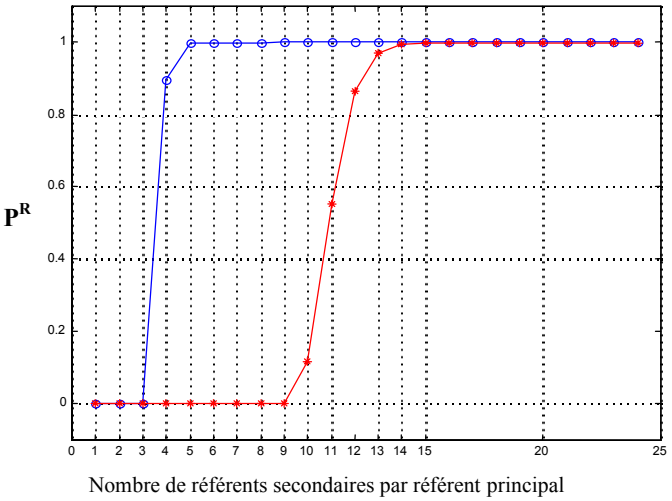

Figure 2. Valeurs de P^R en fonction du nombre r de référents secondaires par référent principal pour 39 langues en moyenne, 34 familles et pour un nombre d'items phonétiques observés (P_{obs}) égal à 2 850 (courbe o), et pour un nombre d'items phonétiques maximal (P_{max})égal à 25 272 (courbe ∗).

Pour que la démonstration de [Ruh97] ait été statistiquement fondée avec les 2 850 items phonétiques contenues dans ses données, la figure 1 montre qu'il aurait fallu ne considérer, au plus, que 2 référents secondaires par référent principal.. En effet, avec 3 référents secondaires, on observe déjà près de 80 % de chances d'obtenir des correspondances et 100 % de chances de trouver des racines communes aux langues répertoriées par un tirage au hasard, même en ne prenant que 4 référents secondaires par référent principal.

5. Discussion

Les données et la méthode utilisées par [Ruh94] [Ruh97] concernant la mise à jour de l'existence de racines universelles est loin d'être convaincante, puisqu'un tirage au hasard des formes phonétiques conduit au même résultat que sa démonstration basée sur des considérations linguistiques (équivalences de formes phonétiques et proximités lexico-sémantiques). Les résultats que nous obtenons amènent à la constatation suivante : avec vingt-quatre référents secondaires (compte tenu du nombre de référents principaux, du nombre de familles linguistiques et du nombre moyen de langues par famille), [Ruh97] avait cent pour cent de chances de trouver, par tirage au hasard, des « racines mondiales » communes à toutes les familles. À partir d'un calcul mathématique et des données de [Ruh94], [Rin96] aboutit à une conclusion similaire. Avec 2 850 formes phonétiques différentes contenues dans les données, il aurait fallu, pour que les résultats qu'il obtient s'écartent de ceux obtenus par tirage au hasard, que [Ruh94] ne considère que trois référents secondaires pour chaque référent principal. Nous montrons que la probabilité est égale à 1 dès cinq référents secondaires pour chaque référent principal. Notre étude montre que Ruhlen [Ruh94] utilise trop peu de référents principaux, beaucoup trop de référents secondaires et pas suffisamment de formes phonétiques différentes relatives au glissement sémantique des racines universelles, même si son choix d'acceptation de changements sémantiques simples élargit les champs lexico-sémantiques et par là même augmente le nombre d'items phonétiques pris en compte dans son analyse.

Cette recherche a été menée dans le cadre du projet OHLL Congruence (Origine de l'Homme, du Langage et des Langues), Resp. Pierre Darlu (INSERM U535).

Bibliographie

[Gre57] Greenberg, Joseph (1957) *Essays in Linguistics*. University of Chicoago Press

[Met01] Métoz, Laurent (2001) *L'hypothèse des « racines mondiales » de Merritt Ruhlen, analyse méthodologique et évaluation statistique*. Mémoire de Maîtrise, Université Stendhal, Grenoble.

[Rin96] Ringe, Donald (1996) The Mathematics of Amerind. *Diachronica* XIII, 135-154.

[Ruh97] Ruhlen, Merritt (1997) *L'Origine des*

[Ruh94] *Langues*. Paris, Débats, Belin (titre original : *The Origin of Language. Tracing the evolution of the mothertongue*. 1994).