INDIAN OCEAN REVIEW OF SCIENCE AND TECHNOLOGY
ISSN: 2312-1874, Article 1-2021

The Sabaki languages of Comoros

Maurizio Serva[1]*, Michele Pasquini[1]

Abstract
We determine by means of quantitative methods the external and the internal cladistics of the four Bantu varieties spoken in the Comoros archipelago. With external cladistics we mean the phylogenetic position of the four varieties with respect to the Sabaki and Makhuwa languages spoken along the South-Eastern coast of Africa. During the years consensus has been reached that Comorian Bantu languages belong to the Sabaki group, but they are different from Swahili, our findings confirm this conclusion. With internal cladistics we mean the mutual classification of these four Comorian languages. Conventionally they are divided into two groups: the Eastern group composed of Shindzwani and Shimaore and the Western group composed of Shimwali and Shingazidja, our results point to a rather different classification with Shingazidja isolated from the other three. Finally, the phylogenetic tree of ten East African languages and the four Comorian ones, which is also constructed, shows a clear bi-partition with Makhuwa languages on one branch and Sabaki languages (included Comorian languages) on the other.

Keywords
Bantu languages — Sabaki languages — Comorian languages — Phylogenesis of languages — Lexicostatistics

[1] *Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy*
***Corresponding author**: serva@univaq.it

Contents

1. Preamble and goals

The Comoros archipelago consists of four islands which lie between the North-West of Madagascar and the South-Eastern coast of Africa. Three of them, Ngazidja (Grande-Comore), Mwali (Mohéli) and Ndzuani (Anjouan), became independent in 1975 and they currently make up a single state entity named Udzima wa Komori (Union des Comores). Maore (Mayotte), on the contrary, chose to remain under French administration, however, on a cultural, ethnical and even religious level, it is part of the same whole as the three other islands.

The Bantu varieties spoken in the archipelago are collectively called Shimasiwa, which means "language of the archipelago" or Shikomori which means "language of the Comoros". Each of the islands has a different variety whose name follows the island name: Shingazidja, Shimwali, Shindzuani and Shimaore (not to mention that each of these varieties is in turn divided into many local dialects).

These four Bantu varieties are the mother languages of the vast majority of the inhabitants of the four islands, the exception being Mayotte where about one third of the natives speaks a Malagasy language (two different varieties) instead of a Bantu language. Nevertheless, although the Malagasy presence in the Comoros dates back at least one millennium [Allibert, 2015], the Malagasy language presently spoken in Mayotte (Kibosy) seems to be the output of recent events, not much older than a pair of centuries [Gueunier, 2004], [Serva and Pasquini, 2021].

In this paper we try to determine the cladistic of these four Bantu languages, both internal and external. With internal cladistic we mean their reciprocal classification, conventionally they are divided into two groups: the Eastern group composed of Shindzwani and Shimaore and the Western group composed of Shimwali and Shingazidja [Rombi, 1984, Rombi, 2003], our findings point to a rather different classification. With external cladistic we intend the phylogenetic position of the Comorian languages with respect the Sabaki and Makhuwa languages spoken along the South-East coast of Africa. During the years consensus has been reached that Comorian Bantu languages belong to the Sabaki group, but they are different from Swahili [Nurse, 1989], [Nurse and Hinnebush, 1993], [Alnet, 2011], our results confirm this conclusion.

2. Dataset and methods

The dataset we use for the research in this paper consists in 14 Swadesh lists [Swadesh, 1952] of 207 items, 4 of which for the Comorian varieties and 10 for the East-African languages (9 from Mozambique and Swahili from Zanzibar). This dataset can be found in Supplementary Material (Dataset S1) A complete overview of the geographical locations of the 14 languages can be appreciated in Fig. 1 where only

the names of the corresponding towns are reported. The link between languages and towns can be found in the captions of the same figure, while the Geodesic coordinates can be found in Supplementary Material (Table S1).

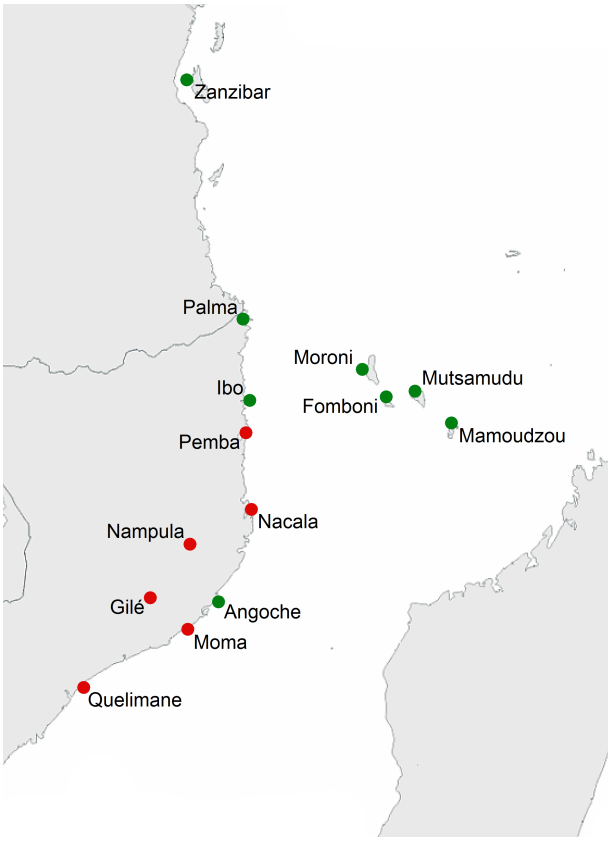

Figure 1. The map with the names of the towns/villages where the 10 East-African languages (nine from Mozambique and one from Tanzania), and the 4 Comorian varieties. The languages and the corresponding towns are Shimwali (Fomboni), Shingazidja (Moroni), Shindzwani (Mutsamudu), Shimaore (Mamoudzou), Swahili (Zanzibar), Swahili (Palma), Mwani (Ibo), Makhuwa-Meetto (Pemba), Makhuwa-Nahara (Nacala), Makhuwa-Central (Nampula), Lomwe (Gilé), Koti (Angoche), Makhuwa-Marrevone (Moma) and Chuwabo (Quelimane). Geodesic coordinates can be found in Supplementary Material (Table S1). The partition is green (Sabaki) versus red (Makhuwa).

The main core of the vocabulary of 207×8 terms of eight Mozambican languages (all the Mozambican languages except the Swahili spoken in Palma) were drawn from the series of dictionaries produced by *SIL Moçambique* (Nampula, Moçambique). These lists where integrated and corrected by Professors Aurelio Simango, Narciso Gastene and Davety Mpiuka of the Universidade Eduardo Mondlane in Maputo with the exception of the lists of Mwani (207 terms), spoken in Capo Delgado, which was integrated by data collected on the field by one of the authors (MS).

The Swahili of Palma (Mozambique) was collected on the field by the same author (MS) while the list of the Swahili of Zanzibar (Tanzania) was drawn up comparing the many available dictionaries.

Finally, the 207×4 terms of the four Comorian languages were taken from the *ORELC* online dictionary http://www.swadrii.com/orelc/ShikomoriWords/?i=kmWords, when possible these terms have been cross-checked by consulting other online dictionaries.

A part of the large amount of information contained in the $N = 14$ short vocabularies which constitute our dataset of about 207×14 words can be encoded into the $N(N-1)/2 = 91$ lexical distances between each pair of languages, which can be considered as the entries of an $N \times N$ upper triangular matrix. Obviously, most of the information is lost in the procedure, but what remains is all what is needed for the goals of this paper.

The lexical distance between two languages is determined by comparing the corresponding two lists of 207 words. In order to carry out this comparison, the present work uses an automated method which was proposed more than ten years ago in [Serva and Petroni, 2008] as a refinement of the historical Swadesh approach [Swadesh, 1952]. In any case, given two languages, say language α and language β, their lexical distance $D(\alpha,\beta)$ is a number between 0 (identical languages) and 1 (totally different languages). The table (matrix) with the $N(N-1)/2 = 91$ lexical distances can be found in Supplementary Material (Table S2),

3. Lexical distances for Comorian languages

The varieties of the Comorian languages are conventionally divided into two groups: the Eastern group with Shindzwani and Shimaore and the Western group with Shimwali and Shingazidja [Rombi, 1984, Rombi, 2003]. This classification goes further in assuming that the differences within each of the two groups are modest. Other classification have been proposed, for example in [Ottenheimer and Ottenheimer, 1976] the Shimaore dialect is placed on one side and the other three varieties on the other.

We would like to add a quantitative point of view to the argument. From Table S2 in Supplementary material, we can extract the following matrix which contains the six lexical distances associated to the six pairs of Comorian varieties:

	Shimwali	Shingazidja	Shindzwani	Shimaore
Shimwali	0.0000	0.2647	0.2032	0.2427
Shingazidja		0.0000	0.3185	0.3604
Shindzwani			0.0000	0.1307
Shimaore				0.0000

We first remark that the largest distances are between Shimaore and Shingazidja (0.3604) and between Shindzwani

and Shingazidja (0.3185). On the other side the smallest distance is between Shimaore and Shindzwani (0.1307). This last distance is so small that we can consider Shimaore and Shindzwani as a single variety (there is, in fact, complete mutual intelligibility). Conversely, Shingazidja remains on the other side, as a different language.

The position of Shimwali can be deduced from the first line of the above matrix. The distance of Shimwali from Shingazidja (0.2647) is not only larger than the distance from Shindzwani (0.2032) but also from Shimaore (0.2427). Therefore, even if Shimwali is somehow intermediate between Shingazidja and Shindzwani/Shimaore, it is closer to the last.

In conclusion, we have the following picture: Shikomori is divided into two groups, one contains only Shingazidja, the other contains the remaining three varieties (two varieties if one considers that Shimaore and Shindzwani are not significantly different one from the other). This picture is at variance with the conventional one and also it is different from the one depicted in [Ottenheimer and Ottenheimer, 1976].

4. Genealogical tree

Once a lexical distance $D(\alpha,\beta)$ of two contemporary languages has been computed from the lists, it is possible to transform it in a genealogical distance $T(\alpha,\beta)$, which is the time from the last common ancestor language. The fundamental formula of Glottochronology states that

$$T(\alpha,\beta) = -\frac{\tau}{2}\ln\left[1 - D(\alpha,\beta)\right],$$

where the value of the characteristic time τ is irrelevant in this paper since we are interested in cladistics, and not in determining the time depth of the Bantu family. For the Malagasy family it was found $\tau = 5136$ years [Serva and Pasquini, 2020], but this value is substantially meaningless when referred to distances involving the Bantu languages since τ is known to be different for different families [Pasquini and Serva, 2019b].

Once computed all the genealogical distances it is possible to construct the genealogical tree, nevertheless, further information is lost in this procedure, especially concerning horizontal transfer (borrowings) [Serva and Pasquini, 2021]. Since in this paper our goal is genealogy, this is not a detrimental, the resulting tree is plotted in Fig. 2 (we have chosen UPGMA, but other choices are equally effective).

As expected, the relative classification of the four Comorian varieties is confirmed: Shikomori is divided into two groups, one contains only Shingazidja, the other contains the remaining three varieties. Moreover, Shimaore and Shindzwani are sufficiently close one from the other to be considered a single variety.

The tree of the 14 languages has two main branches, the green one corresponds to Sabaki group, while the red one to the Makhuwa group, which explain the colours in Fig. 1. The Comorian foursome clearly belongs to the Sabaki branch, but it is well genealogically separated from Swahili which is a different language.

As it is well known, Mwani also belongs to the Sabaki group and its position in the green branch is not surprising. Koti also lies in the green branch, and our point is that it also belongs to the Sabaki group. This statement is debated, Koti is sometimes classified as a Makhuwa languages deeply influenced by Swahili and other times classified as a Sabaki language deeply influenced by Makhuwa.

We don't think that the position of Koti on the green branch is due to a discrepancy between the true phylogeny and a biased tree generated by lexical distances whose values are deeply influenced by borrowings. In fact, genealogy is usually unveiled by standard tools based on vocabularies even in presence of strong contamination. Consider, for example, the place of English in the genealogical tree of Indo-European languages in [Serva and Petroni, 2008]: despite the relatively small distance from French, English is rightly placed in the Germanic branch.

5. Summary and outlook

Our results can be summarized as follows: Comorian is divided into two groups, one contains only Shingazidja, the other contains Shimwali, Shindzwani and Shimaore. Moreover, Shimaore and Shindzwani are not significantly different one from the other to be considered as different varieties. For what concerns the position of the Comorian foursome among the Bantu family, we confirm that is belongs to the Sabaki group, while being undoubtedly distinct from Swahili.

Although our findings may have an intrinsic interest, we consider this study as a preliminary research which aims to clarify a still controversial aspect of the history of the peopling of Madagascar which concerns the role of Africa and the Bantu populations. In linguistics the prevalent belief is that the Indonesian sailors settled some spots on the Eastern African coast before moving to Madagascar [Adelaar, 2012, Adelaar, 2021]. This African stopover allowed for a large genetic horizontal transfer from Africans to Indonesians and also allowed for a small amount of lexical horizontal transfer. In genetics, on the contrary, the prevalent hypothesis is that the Bantu settling followed the Indonesian colonization [Cox *et al*, 2012, Pierron *et al*, 2017, Heiske *et al*, 2021], which implicitly entails a direct navigation from Indonesia to Madagascar, even if a role of East Africa in the Austronesian expansion is not denied [Brucato *et al*, 2019].

In [Serva and Pasquini, 2020] it was provided a strong argument, based on the notion of diversity in linguistics, which favours a South-East Madagascar landing of the Austronesian colonizers, while arguments concerning oceanic currents can also add further clues [**?**] to this location for the first arrival. In turn, a South-Eastern disembarkation is compatible with a direct navigation from Indonesia.

We think that a deeper look to the lexical horizontal transfer between Eastern Africa, Comoros and Madagascar could

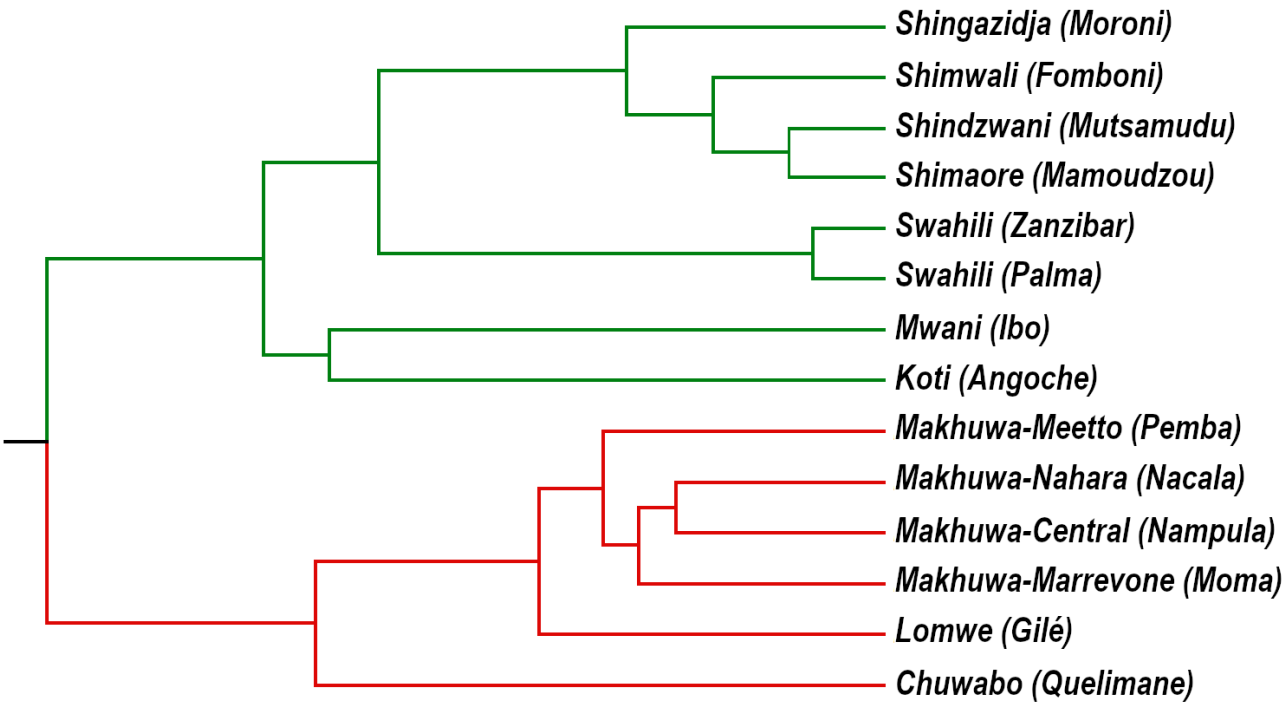

Figure 2. The UPGMA tree of the 14 languages. There are two main branches, the green one corresponds to Sabaki group, while the red one to the Makhuwa group, which explain the colours in Fig. 1. Shingazidja turns out to be separated from the remaining three varieties. Shimaore and Shindzwani are sufficiently close one from the other to be considered a single variety. The Comorian foursome clearly belongs to the Sabaki branch, but it is well genealogically separated from Swahili which is a different language. Koti turns out to be a Sabaki language, probably the southernmost of the group.

be conclusive for a solution of this controversy.

Acknowledgments

We are grateful to the Professors Aurelio Simango, Narciso Gastene and Davety Mpiuka for their invaluable advice concerning the nine Mozambican languages used in the present study.

Supplementary material

Dataset S1: *Comoros and East Africa Coast Swadesh lists.* The complete dataset of the 14 Swadesh lists (four from Comoros, nine from Mozambique and one from Tanzania).

Table S1: *Languages, towns and coordinates.* The names of the 14 languages and the names of the corresponding Towns with their geodesic coordinates.

Table S2: *Lexical distances.* The matrix with the N×(N–1)/2 = 91 lexical distances between each pair of languages.

References

[Adelaar, 2012] A. Adelaar, *Malagasy Phonological History and Bantu Influence.* Oceanic Linguistics **51**, 123-159, (2012)

[Adelaar, 2021] A. Adelaar, *Language contact in Africa.* To appear in *The Oxford Guide to the Malayo-Polynesian languages of South East Asia,* A. Adelaar and A. Schapper editors. Oxford Guides to the World's Languages, Oxford: OUP, (2021).

[Allibert, 2015] C. Allibert, *L'archipel des Comores et son histoire ancienne. Essai de mise en perspective des chroniques, de la tradition orale et des typologies de céramiques locales et d'importation.* Afriques **06**, L'Afrique orientale et l'océan Indien: connexions, réseaux d'échanges et globalisation, (2015).

[Alnet, 2011] A. Johansen Alnet, *The clause structure of the Shimaore dialect of Comorian (Bantu).* ProQuest, UMI Dissertation Publishing, (2011).

[Brucato *et al*, 2019] N. Brucato, V. Fernandes, P. Kusuma, V. Černý, C. J. Mulligan, P. Soares, T. Rito, C. Besse, A. Boland, J.-F. Deleuze, M. P. Cox, H. Sudoyo , M. Stoneking, L. Pereira and F.-X. Ricaut, *Evidence of Austronesian Genetic Lineages in East Africa and South Arabia: Com-*

plex Dispersal from Madagascar and SouthEast Asia. Genome Biology and Evolution **11**, 748-758, (2019).

[Cox *et al*, 2012] M. P. Cox, M. G. Nelson, M. K. Tumonggor, F-X. Ricaut and H. Sudoyo, *A small cohort of Island SouthEast Asian women founded Madagascar* Proceedings of the Royal Society B **279**, 2761-2768, (2012).

[Gueunier, 2004] N. J. Gueunier, *Le dialecte malgache de Mayotte (Comores): une discussion dialectologique et sociolinguistique.* Faits de Langues, **23**, 397-420, (2009).

[Heiske *et al*, 2021] M. Heiske, O. Alva, V. Pereda-Loth, M. Van Schalkwyk, C. Radimilahy, T. Letellier, J.-A. Rakotarisoa and D. Pierron, *Genetic evidence and historical theories of the Asian and African origins of the present Malagasy population* . Human Molecular Genetics, https://doi.org/10.1093/hmg/ddab018, to appear, (2021).

[Nurse, 1989] D. Nurse and T. J. Hinnebush, *Is Comorian Swahili? Being an examination of the diachronic relationship between Comorian and coastal Swahili.* In *Le Swahili et ses limits.* M.-F. Rombi editor, Paris, (1989).

[Nurse and Hinnebush, 1993] D. Nurse and T. J. Hinnebush, *Swahili and Sabaki: A linguistic history.* Berkeley: University of California Press, (1993).

[Ottenheimer and Ottenheimer, 1976] H. Ottenheimer and M. Ottenheimer, *The classification of the languages of the Comoros islands.* Anthropological Linguistics, **18**, 408-415, (1976).

[Pasquini and Serva, 2019b] M. Pasquini and M. Serva, *Stability of meanings versus rate of replacement of words: an experimental test.* Journal of Quantitative Linguistics , (online 6 Aug 2019, DOI: 10.1080/09296174.2019.1647754).

[Pierron *et al*, 2017] D. Pierron, M. Heiske, H. Razafindrazaka, I. Rakoto, N. Rabetokotany, B. Ravololomanga, L. M.-A. Rakotozafy, M. Mialy Rakotomalala, M.Razafiarivony, B. Rasoarifetra, M. Andriamampianina Raharijesy, L. Razafindralambo, Ramilisonina, F. Fanony, S. Lejamble, O. Thomas, A. M. Abdallah, C. Rocher, A. Arachiche, L.Tonaso, V. Pereda-loth, S. Schiavinato, N.Brucato, F.-X. Ricaut, P. Kusuma, H. Sudoyo., S. Ni, A. Boland., J.-F. Deleuze, Ph. Beaujard, Ph. Grange, S. Adelaar, M. Stoneking, J.-A. Rakotoarisoa, C. Radimilahy, and T. Letellier, *Genomic landscape of human Diversity across Madagascar.* Proceedings of the National Academy of Sciences **114**, E6498-E6506, (2017).

[Rombi, 1984] M.-F. Rombi, *Le shimaore, première approche d'un parler de la langue comorienne* Langues et Cultures Africaines **3**, Paris, SELAF, (1984).

[Rombi, 2003] M.-F. Rombi, *Les Langues de Mayotte.* In *Les langues de France*, M. Alessio and J. Sibille editors (dir. B. Cerquiglini), PUF, Paris, 305-318, (2003).

[Serva and Petroni, 2008] M. Serva and F. Petroni, *Indo-European languages tree by Levenshtein distance.* EPL **81**, 68005, (2008).

[Serva and Pasquini, 2020] M. Serva and M. Pasquini, *Dialects of Madagascar.* PLoS ONE **5(10)**, e0240170, (2020).

[Serva and Pasquini, 2021] M. Serva and M. Pasquini, *Malagasy dialects in Mayotte.* EPL **133**, 68003, (2021).

[Swadesh, 1952] M. Swadesh, *Lexicostatistic dating of prehistoric ethnic contacts.* Proceedings of the American Philosophical Society **96**, 452-463, (1952).