A Computational Analysis of Gabon Varieties

Bart Alewijnse, John Nerbonne
Humanities Computing
University of Groningen
B.Alewijnse@student.rug.nl,
J.Nerbonne@rug.nl

Lolke J. van der Veen
Dynamique du Langage (UMR 5596)
CNRS, Université Lumière-Lyon 2
Lolke.van-der-Veen@univ-lyon2.fr

Franz Manni
Musée de l'Homme MNHN, Paris
manni@mnhn.fr

Abstract

The linguistic situation in Gabon is highly complex as the various varieties form long chains, and multilingualism is common. Most previous classifications of Gabon varieties have used lexical models, a notable exception being Nurse & Philippson (2003), which involves phonological and morphological features.

This paper presents a phonetic analysis of Gabon varieties. It applies Levenshtein analysis to obtain pronunciation distances, which are in turn analyzed using bootstrapped clustering to identify groups, including an estimation of the robustness of the clusters. On the basis of clustering, we obtain a graded map of varieties.

The results indicate that techniques developed and proven on European languages are still useful when applied to Bantu in spite of its different structure and its nomadic speakers.

Keywords

Levenshtein distance, Bantu

1 Introduction

The present paper applies a measure of pronunciation distance to Gabon Bantu varieties in an effort to detect their relatedness.

Gabon is located in western central Africa, bordered by Cameroon on the north, by Equatorial Guinea to the northwest, by Congo to the east and south, and by on the west by the Atlantic ocean. See Fig. 1. It has a population of approximately 1.2 million. The official language is French, and its population collectively speaks over fifty local language varieties, all of which are Bantu with the exception of Baka, a Pygmy language, which is Ubangian.

The phonetic measure is a variant of EDIT DISTANCE or LEVENSHTEIN DISTANCE, and this is the first application of this sort of analysis to Bantu language variants, or indeed any African languages for the purpose of detecting linguistic affinities. A general purpose of this paper is therefore to verify that the techniques developed for European languages can successfully be applied to Bantu. A longer term goal that will not be realized in this paper is to compare linguistic and extralinguistic measures of relatedness, in particular, genetic relatedness.

The Gabon Bantu varieties are particularly interesting with respect to the edit distance measure due to their extensive use of prefixation, which has the potential to confuse the alignment in the edit distance measure. The Gabon Bantu varieties are also geographically interesting as the Bantu people have been unusually mobile, disrupting the usual geographic cohesion of language variation. Gabon Bantu speakers are also known to mix languages extensively, another potential challenge for the techniques.

The phonetic data being studied comes from an as of yet unpublished database, which will be introduced briefly.

2 Previous work

The Bantu language varieties in Gabon are classified as part of Western Bantu and Forest Bantu (which according to Nurse & Philippson (2003), henceforth: N&Ph (2003), is a subset of Western Bantu).[1] They belong to Guthrie's zones A, B and H.

Note that Maho (2003) proposes an update of Guthrie's classification, which attributes new codes to missing languages. These codes are easily distinguishable from the ones used in Guthrie's original list since we refer to varieties not listed in Maho (2003) using Guthrie's codes followed by the locations of specific varieties between parentheses.

A northwestern vs. (central-)western split within Western Bantu is well supported [1, 6, 10], with zone A language variants + B10 (Myene), B20 (Kele), and B30 (Tsogo) considered as part of northwestern Bantu, and the remaining language varieties as part of (central-)western Bantu. N&Ph (2003) do not provide direct evidence in favor of or against such a split, as they did not examine higher level groupings.

Local, lower-level clusters may be identified (sometimes transcending the current borders) such as [A75], [A80], [B10 and B30], [B20 (?B21)], [B40-some H12 and H13], [B50 and B73], [B60] and [B70 (less B73, B81, B83-4)], as well as intermediate groupings such as [B10-30], [B50-60-70, parts of B80-H24] and [H10,

[1] Note N&Ph (2003)'s definition of Western Bantu is not necessarily the same as other definitions; cf. Grégoire (2003)

H30, H42, B40-parts of B85] (cf. N&Ph (2003)). See Fig. 2 for the geographic locations of the sites. However, the identification of higher-level entities is particularly arduous. New criteria emerging from the study of verb morphology may allow researchers to tackle this issue more effectively.

3 Data

3.1 Summary

The data used in this study has not been analyzed before and is part of a dataset under development for the *Atlas Linguistique du Gabon* (ALGAB), a database planned for release in 2010. The data being studied will be made available via the *it Dynamique du Langage* website,[2]

The data can be presented in a table representing phonetic data points for 160 glosses (concepts), at 53 sampling sites. The table is an aggregation of various samplings at different places and times, and somewhat sparse for various reasons.

As a rule, both singular and plural forms have been collected, though for some varieties there is only one form. Having singular and plural forms is important to Bantu specialists for morphological information such as finding the gender of substantives which is reflected by the choice of plural prefixes.

Although tone and stress information have been ignored in this study, the authors do not assume these features to be less relevant, only that both require a fuller treatment after more careful study.

Stress is not marked in the database because it is predictable in all varieties. It is usually placed systematically on the first syllable on the noun stem, while sometimes straightforward penultimate stress is used. No stress contrasts have been found (within single varieties). While the decision not to mark stress is understandable from the point of phonological theory, we would prefer to have data marked with stress to keep track of its distinctive use among different varieties.

As far as we know, tone is indeed distinctive in most if not all varieties. Previous analysis has revealed a few different basic categories of tone systems in use, which is one among several details that make proper study and verification of tonal transcription throughout all the data very time consuming. Since tone has not been systematically transcribed in the field (for different reasons, including absence of tonal contrast at the surface, and because of the priority given to the segmental level, or due to the ability of the consultant), tonal information had to be discarded from the data in this analysis for ease of comparison.

The table of data has 10417 filled cells, approximately 64% of the possible whole. There are a few more data points as some entries consist of more than one linguistic equivalent.

There are two relatively frequent diacritics present in the data, nasalization and the syllabic marker.

[2] See `http://www.ddl.ish-lyon.cnrs.fr/`. ALGAB has been used in a few other papers [15, 7, 11, 9], including PhD theses and local working papers which can be found on the DDL website under author names like Hombert, Blanchon, Fontaney, Mouguiama-Daouda, Van der Veen, and others.

3.2 Collection Objectives, Locations, Time Span

As the overall linguistic situation of Gabon was rather poorly understood in the early 1980s, a small team of Africanists working in Lyon decided to launch an extensive language survey. This carefully planned and organized survey led to the discovery of several unknown varieties (some of which are extinct by now), and to a deeper understanding of the local languages and the relationships between them.

The team surveyed province by province. Traveling was done by car, by pirogue or on foot, from one village to another, following the main axes of the country (roads, paths, and rivers). Libreville, where one can find speakers of virtually all of the languages of the country, has become an important place for retrieving possible missing links, establishing new contacts, and completing the survey. Fig. 2 shows the locations of the sites where data was collected, and the appendix provides village names and Ethnologue labels to allow identification.

Data was collected in the field during several short-term missions in two major periods, 1985-1991 and 2000-2005, but from 1990 on also in Lyon and Tervuren (Belgium), mainly by postgraduate students. Different provinces have been sampled at different times.

Data collection was an essential part of a preliminary linguistic inquiry with classification and description in mind, including the elaboration and publication of a linguistic atlas of the Gabon area, the still ongoing ALGAB project that started in the 1980s), the study of basic phonology and morphology (nouns, verbs), and a series of preliminary comparative and diachronic studies (reflexes of the proto-language, regional reconstructions, borrowings). These studies have resulted in a considerable number of publications and dissertations (MA, PhD). See above, note 2.

3.3 Sample construction and field work

The ALGAB word list was designed for preliminary linguistic research depending on the linguistic and cultural situation of Gabon. It draws on existing elicitation lists such as the ALAC list[3] and takes previous experience and knowledge of the (extended) area into account.

The list of 160 words includes mainly nouns (89) and verbs (41), and additionally numerals (from one to ten), adjectives (13), adpositions (2), interrogative pronouns (2) and a few unclassifiable items. The set was chosen to obtain high-frequency core vocabulary that is not culturally marked, at least not to a great degree.

Fieldwork was performed by a team comprising some 15 well-trained elicitors: Jean-Marie Hombert, Gilbert Puech, Jean Alain Blanchon, Louise Fontaney, Lolke Van der Veen, Pither Medjo Mvé, Patrick Mouguiama-Daouda, Daniel-Franck Idiata and Roger Mickala-Manfoumbi. The few initial and principal elicitors (Hombert, Puech, Blanchon, Fontaney) are all

[3] Atlas Linguistique de l'Afrique Centrale. See also Dieu and Renauld (1983).

experienced fieldworkers and worked closely with less experienced, participants, often supervising them.

Consultants were chosen in various ways. Whenever possible, the choice was made in consultation with the elders of the communities or, failing that, based on a preliminary check. Elicitation was usually carried out in French with a bilingual speaker, while in a few cases through an interpreter.

Many interviews were conducted in villages and hamlets, but others took place more informally on the roadside. In most cases, several speakers have been interviewed for each of the language varieties.

Both data collection activities and the data itself were documented carefully, including as many details as possible: language varieties with their name(s), dates, names of consultants, names of elicitors, number of items collected, nature and quality of elicited material, locations, maps with precise or approximate location(s) for each language variety, etc.

The collected data was subsequently checked systematically with the help of additional consultants and using good quality recordings made in the field (as a rule, word lists were recorded in the field using DAT recorders or mini-disk recorders). The sound recordings were particularly important in checking transcriptions by less experienced elicitors, where they served to safeguard the uniformity and the reliability of the data. Additionally, judgments of reliability were attributed to each sample collected in the field, which resulted in some data being discarded. Overall, the data was thoroughly checked.

Sample lists may incomplete for several reasons. Many of the varieties of Gabon are nearly extinct, and their speakers are not always able to recall the equivalents of the entries of the word list. In addition, multilingualism being the rule, speakers tend to mix up languages. In several cases, lists are incomplete because of a lack of time. This also explains why certain samples merely contain the initial, i.e. noun, part. Since the task of a language assistant is tedious, another understandable reason is a lack of motivation on behalf of the consultants, who all participated on a voluntary basis.

3.4 Transcription

The data used for this analysis is a careful simplification of a larger database under development in Lyon. This version was transformed based on an up-to-date analysis of the respective language variants; predictable features such as contextual nasalization or lengthening have not been retained.

3.5 Representation, Conversion

The data was supplied in a Unicode encoding, but not in Unicode IPA, rather in an encoding which uses a special set of characters which must be viewed in combination with the IPALA font. Conversion to a more standard format was therefore necessary before analysis. Since our current models are implemented using X-SAMPA, the IPALA-coded characters were mapped to X-SAMPA. This conversion was verified, since IPALA is not fully documented.

Table 1 shows the resultant phonetic characters, as IPA and X-SAMPA characters, together with their frequency distribution.

X-SAMPA	IPA	occurrences			
@	ə	948			
1	ɨ	7			
E	ɛ	1505			
O	ɔ	1950	b	b	2809
o	o	2650	d	d	1814
a	a	8248	g	g	1102
e	e	2139	f	f	303
i	i	5655	h	h	145
I	ɪ	42	k	k	2197
u	u	4489	j	j	1482
U	ʊ	71	m	m	4505
V	ʌ	1	l	l	2224
Q	ɒ	1	n	n	2484
G\	ɢ	2	p	p	764
4	ɾ	3	s	s	1205
?	ʔ	19	r	r	554
p\	ɸ	2	t	t	2191
B	β	226	w	w	1119
D	ð	18	v	v	264
G	ɣ	970	x	x	13
H	ɥ	2	z	z	532
J	ɲ	642	_~	̃	40
N	ŋ	1325	_=	ˌ	95
S	ʃ	520			
R	ʁ	27			
T	θ	3			
Z	ʒ	424			

Table 1: Phonetic characters as X-SAMPA and IPA, together with their frequencies in the Gabon data set.

3.6 Examples

To give a rough illustration of the phonetic detail and variation between varieties, we provide a small excerpt of the phonetic data, seven words at two sites, one in each Guthrie zone.

	A34	**B42 (Mimongo)** singular	**B42 (Mimongo)** plural
animal	tito	ɲamə	baɲamə
fat, oil	iβoŋgo	maatsi	maatsi
intestine	miya	musopu	misopu
nose	βiho	mbasu	bambasu
rope	ukɔdi	mukudu	mikudu
wind	upupɛ	diβuɣə	maβuɣə
woman	mwadyo	muɣyetu	baɣyetu

The word for 'rope' illustrates the problem of prefixes and the influence the prefixes have on pronunciation. We further note that it would be surprising to find this degree of variation in the dialect atlas of a contemporary European language.

3.7 Geographic data

The data represents sites spread throughout much of Gabon, sometimes in close proximity, and in two cases across the border in the Congo. Exact coordinates of

many collection sites were provided, while other locations were only described. Gazetteers were used to verify and augment the list as much as possible. A few locations were calculated from fairly detailed descriptions such as "75km north of Z" or "between X and Y", where X and Y were fairly close.

Other location names or descriptions refer approximately to a collection site, or have a name that refers to one of several sites in gazetteer data, usually related ones. Because of this a number of locations are not exact, namely B11a, B11d, B22b, B20x, B31, B32, B304, B42, B252, B305, B602, B71a (Ossele), B71a (Ibali) and B71a (Djoko), which are shown on the first map with hollow markers (Fig. 2)

But the vagueness in the reference of place names is not the only problem in locating the provenance of linguistic varieties. In addition, respondents were not always sure where their group or tribe was normally located, *inter alia* because the members had moved a good deal, and because several varieties are scattered rather widely. Taken together, these problems mean that we should exercise caution in reasoning about the influence of geography.

4 Techniques used

The L04 dialectometric package,[4] developed at Groningen university, was used for the bulk of the calculations.

We note that missing values are basically ignored in analysis: we calculate the distance between two sites based on the pronunciations present and calculate the mean distance for all the words that are compared. This means that some dialect distances are based on more comparisons than others and are therefore more reliable statistically, but there is a large amount of data, so that no comparisons are unreliable.

In the present study some varieties record singular and plural forms for each gloss, while others have others a single form. This would make location comparison nontrivial, but L04 handles this inequality by seeking optimal matches and uses the mean of those. In the cases where one variety has one form and the other two, the comparison boils down to the average of the two distances.

4.1 Levenshtein distance

We compared pronunciations using Levenshtein distance, which may be understood as the cost of the optimal set of operations need to map one string to another. Heeringa (2004) contains an extensive introduction to the application of Levenshtein distance to the problem of measuring the distance between pronunciations.

The phonetic model has discrete costs, meaning that identical tokens cost nothing, while vowel-vowel and consonant-consonant substitutions cost one unit, as do insertions and deletions. In general this version of the algorithm only allows substitutions respecting SYLLABICITY, i.e. vowels for vowels and consonants for consonants. There are three exceptions to strict vowel-consonant borders: the semivowels [j] and [w] as well as the maximally high vowels [i] and [u] may match both vowels and consonants, and [ə] may match sonorant consonants.

Consonant-vowel substitutions are much more expensive than the combination of a deletion and insertion to the same effect, which enforces the syllabicity constraint, and also causes the Levenshtein results to have slightly longer alignments that are usually more natural.

Diacritics are not considered by the present model, meaning that the ninety-five occurrences of syllabic markers (marking syllabic sonorants) and the forty occurrences of nasalization are ignored. These counts are low enough with respect to the overall sample site so that we are confident that results were not affected greatly.

Following the analysis of [5], a model was used that attempts to respect phonetic context by applying the phonetic model not to words represented as sequences of character unigrams, but rather to words represented as sequences of character bigrams, thereby including effects of (direct) phonetic context. The resulting comparison costs were not normalized by length, also following Heeringa et al.'s (2006) findings.

The result of the pairwise distance measures between all sites is a difference matrix containing linguistic distances between all pairs of sites. Cronbach's α is calculated as measure of consistency in the data, and was determined to be a nearly perfect 0.93 (based on the full dataset), meaning that we have enough data for a clear signal, while the correlation between the linguistic distances and the geographic distances was calculated to be 0.461. We interpret the latter to mean that geography clearly influences Bantu linguistic similarity in Gabon, but not overwhelmingly.

5 Results

5.1 Line map

The line map (Fig. 3(a)) visualizes the distances between all site pairs. This figure shows the mean phonetic distances between each site without any further processing, and so reflects the results of Levenshtein analysis transparently, but is visually rather dense and does not clearly reveal groups (for example in cases of sites near each other).

5.2 Clustering

We employ bootstrap clustering in order to identify stable groups in the data. We use a bootstrap procedure because hierarchical agglomerative clustering is not in general stable—small changes in input data can change the "minima" that are sought in clustering, leading eventually to large changes in the groupings found. This also means, however, that the procedure may be sensitive to noise.

To overcome the problem of instability, we apply a bootstrapping step that can be described roughly as using clustering repeatedly, using many random selections of the data (selecting with replacement). The entire collection of clusterings is then inspected to see

[4] `http://www.let.rug.nl/~kleiweg/L04/`

Figure 1: Location of Gabon

Figure 2: Sampling locations in Gabon. Empty circles indicate approximate locations. See the Appendix for a list of village names and Ethnologue labels.

which groups emerge reliably. We use UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clusterings although we have experimented with others [13].

We present the consensus dendrogram in Fig. 4(a). The numbers associated with the groups indicate how reliably that group emerged, from a total of 100 runs. For example, clusters where '60' is adjoined were found in 60 of 100 runs, but not in the other forty. The evidence for these clusters is much less reliable that that for clusters found more than 90 times. The length of the branches in the dendrograms indicate the mean distance at which the groups were found, i.e. the so-called COPHENETIC DISTANCE. We shall interpret these distances in subsequent processing.

5.3 Multi-Dimensional Scaling

The bootstrapped difference matrix was also analyzed via Multi-Dimensional Scaling (MDS), a dimensionality reduction algorithm. The use of MDS in other dialectological applications has allowed us to visualize the notion of a dialect continuum in a well-founded way [12]. Normally we apply MDS directly to dialect distance matrices, but here we apply it to the mean cophenetic distances which result from bootstrap clustering. The result is shown in Fig. 3(b). This is a novel sort of visualization, which we have not been able to present at length. It tends to emphasize the effect of clustering.

When we apply MDS to the mean cophenetic distances from the consensus (bootstrap) clustering, we find a good correlation between the original distances and the distance in the two-dimensional MDS ($r = 0.697$), and a slightly better result in three dimensions ($r = 0.762$). These figures indicate the amount of dialect variation that may be explained in models of this reduced form.[5] We note that the third dimension reduces the unaccounted for variance by $0.08 = (0.76^2 - 0.70^2)$, which is a 16% reduction. The three-dimensional data was used in a reduced dimensionality map, which uses the three dimensions as color components using the RGB color model.

6 Discussion of results

We conclude that the techniques we have applied to Indo-European languages in earlier work may also be applied to Bantu languages. The special linguistic features of the languages did not present an insurmountable problem. The mobility of the Gabon Bantu population has meant that we need to refer more to the consensus dendrogram analysis of data than to the maps displaying the results. We turn now to the specifics of the affinities we noted.

Within the expected complex network shown in Fig. 3(a), many lower-level groups appear that match our expectations perfectly: [A75], [B10], [most of B20], [B30], [B40], etc. B21's well-known isolated position also clearly appears here (and there is no linguistic proximity to A34 (Benga), although they end up near each other in the dendrogram).

Though B20 forms the most scattered group within territory surveyed, its presumed members do group very reliably (cf. Tombidi (B20x) and B201 (Ndasa) in the south), which suggest some definite underlying unity, with the exception of B24 and B21. The faint link between A34 and B25 (Kota) should also attract our attention: it corroborates both earlier (unpublished) linguistic studies and oral tradition (not visible in the MDS plot, Fig. 4(b)). There is no evidence of a link between B11a (Mpongwe) and A34, although the latter is clearly dominated by the former nowadays.

As far as the bootstrapped clustering (Fig. 3(b)) is concerned, two northern groups appear: A75 (the Fang dialect cluster), and part of B20 (B20x=Mwesa; B22b, B252, B25). With respect to the latter, links appear with various other areas with yellow and yellowish shades, especially in the south and in the surroundings of Lambaréné (B22). The different yellow shades suggest the unity of B20, with varying degrees of internal distances. This unity has been questioned by some scholars and has never been proven. Bastin et al. (1999), have found B20 to be a floating group, clustering with northern languages in some cases and with southern languages in other.

The collection sites in the southwest (colored in lilac) perfectly match the SHIRA group (B40).

As expected, B10 and B30 varieties cluster together (sites colored in red). B32 (Okande) correctly clusters with the other B30 varieties, in spite of its geographical eccentricity.

In the line map (Fig. 3(a)) the B30 and B10 groups appear to form a group, i.e. a sort of ("central belt"), which corroborates previous observations to this effect, although the reason for this apparent convergence is still a matter of debate. However, as inspection of the consensus dendrogram and the MDS-reduced map of mean cophenetic distances confirm, this is overemphasized because some of the relevant data points are so close together that they are difficult to distinguish visually.

7 Future work

Future steps in this analysis should include the extraction of the dominant linguistic sources of the aggregate differences, a more detailed comparison to existing scholarly literature, and, hopefully, the opportunity to compare the linguistic landscape with the distribution of other traces of cultural similarity and population history.

8 Credits

Thanks go to the data's many collectors, and to Peter Kleiweg, who wrote and maintains L04, the dialectometric package used for much of the processing. This investigation been performed in the context of a (planned) comparison of genetic and linguistic variety, as part of the OHLL[6] program of the French

[5] The percent of explained variance is $100 \times r$

[6] *Origine de l'Homme, du Langage et des Langues*

CRNS[7], "*Contribution à l'étude des langues bantoues et des peuples bantouophones: approche linguistique, approche génétique*", and also in extension of the ESF Eurocores OMLL[8] program "Language, Culture, and Genes in Bantu: A Multidisciplinary Approach", both coordinated by L. J. Van der Veen (UMR 5596 *Dynamique du Langage*, Lyon).

References

[1] Y. Bastin, A. Coupez, and M. Mann. Continuity and divergence in the Bantu languages: Perspectives from a lexicostatistic study. In *Annales des sciences humaines, vol. 162.* Musée royal d'Afrique centrale, Tervuren, 1999.

[2] M. Dieu and P. Renaud. Situation linguistique en afrique centrale : Inventaire préliminaire, le cameroun. In *Atlas linguistique de l'Afrique Centrale (ALAC); Atlas linguistique du Cameroun (ALCAM).* Paris-Yaounde, 1983.

[3] C. Grégoire. The Bantu languages of the forest. In G. Nurse, & Philippson, editor, *The Bantu Languages*, pages 349–370. Routledge Language Family Series, London/New York, 2003.

[4] W. Heeringa. *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* PhD thesis, Rijksuniversiteit Groningen, 2004.

[5] W. Heeringa, P. Kleiweg, C. Gooskens, and J. Nerbonne. Evaluation of string distance algorithms for dialectology. In *Proceedings of the Workshop on Linguistic Distances*, pages 51–62, Sydney, Australia, July 2006. Association for Computational Linguistics.

[6] C. Holden and R. Gray. Rapid radiation, borrowing and dialect continua in the Bantu languages. In P. Forster and C. Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, Cambridge, 2006. The MacDonald Institute for Archaeological Research.

[7] D. F. Idiata-Mayombo. *Aperçu sur la morphosyntaxe de la langue isangu (Bantu, B42).* PhD thesis, Lincom Studies in African Linguistics, 32. Munchen/Newcastle, 1998.

[8] J. Maho. A classification of the Bantu languages: an update of guthrie's referential system. In D. Nurse and G. Philippson, editors, *The Bantu Languages*, pages 639–651, London/New York, 2003. Routledge Language Family Series.

[9] P. Mouguiama-Daouda. *Les dénominations ethnoichtyologiques chez les Bantous du Gabon: étude de linguistique historique.* PhD thesis, Université Lumière-Lyon 2, 1995.

[10] P. Mouguiama-Daouda. Contribution de la linguistique à l'histoire des peuples du gabon, la méthode comparative et son applications au bantu. In *Collection Sciences du Langage*, Paris, 2005. CNRS Editions.

[11] P. M. Mvé. *Essai sur la phonologie panchronique des parlers fang du Gabon et ses implications historiques.* PhD thesis, Université Lumière-Lyon 2, 1997.

[12] J. Nerbonne, W. Heeringa, and P. Kleiweg. Edit distance and dialect proximity. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, pages pp.v–xv., Stanford, 1999. CSLI Press.

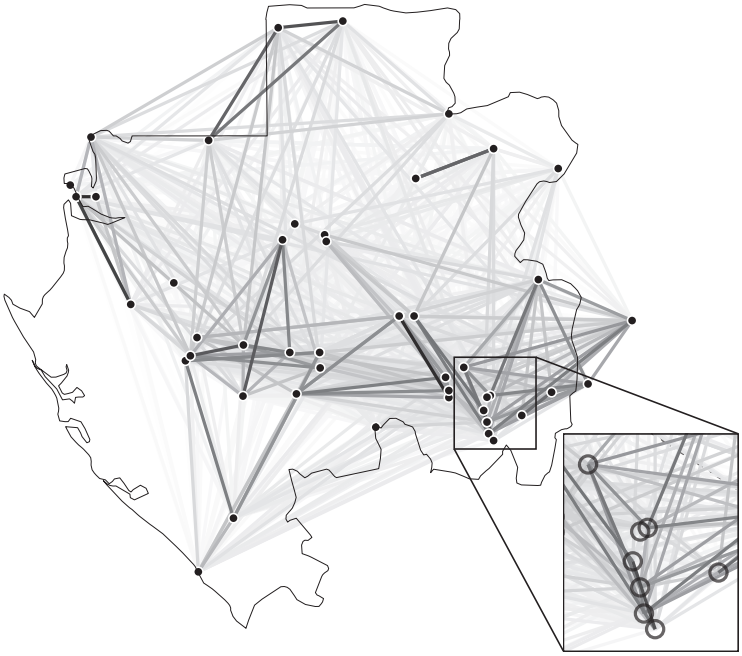
[13] J. Nerbonne and P. Kleiweg. Toward a dialectological yardstick. *Quantitative Linguistics*, 14(2):148–167, 2007.

[14] D. Nurse and G. Philippson. Towards a historical classification of Bantu languages. In D. Nurse and G. Philippson, editors, *The Bantu Languages*, pages 164–181. Routledge Language Family Series, 2003.

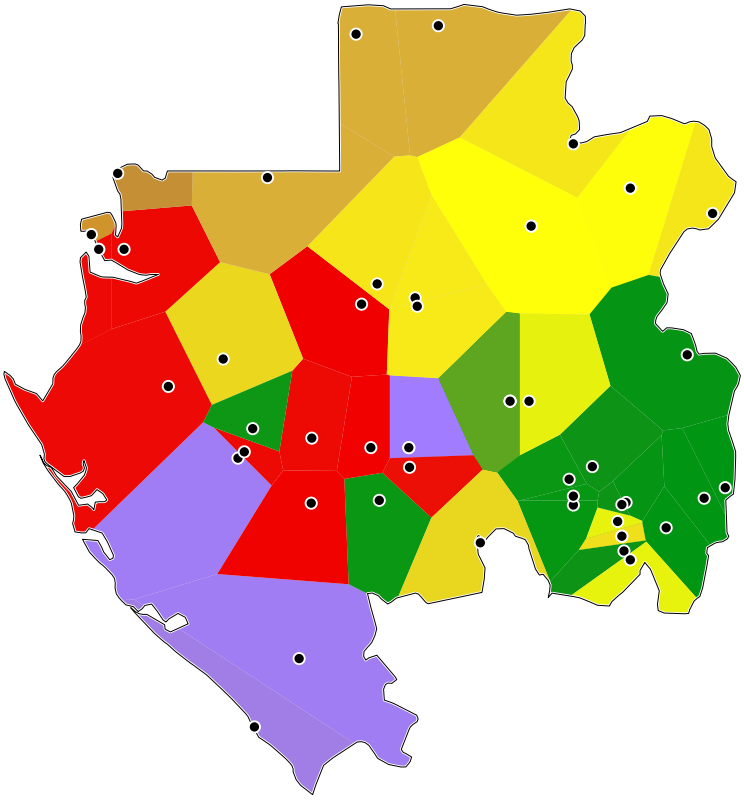
[15] L. J. van der Veen. *Etude comparée des parlers du groupe Okani, B30 (Gabon).* PhD thesis, Université Lumière-Lyon 2, 1991.

[7] *Centre National de la Recherche Scientifique*

[8] Origin of Man, of Language and of Languages

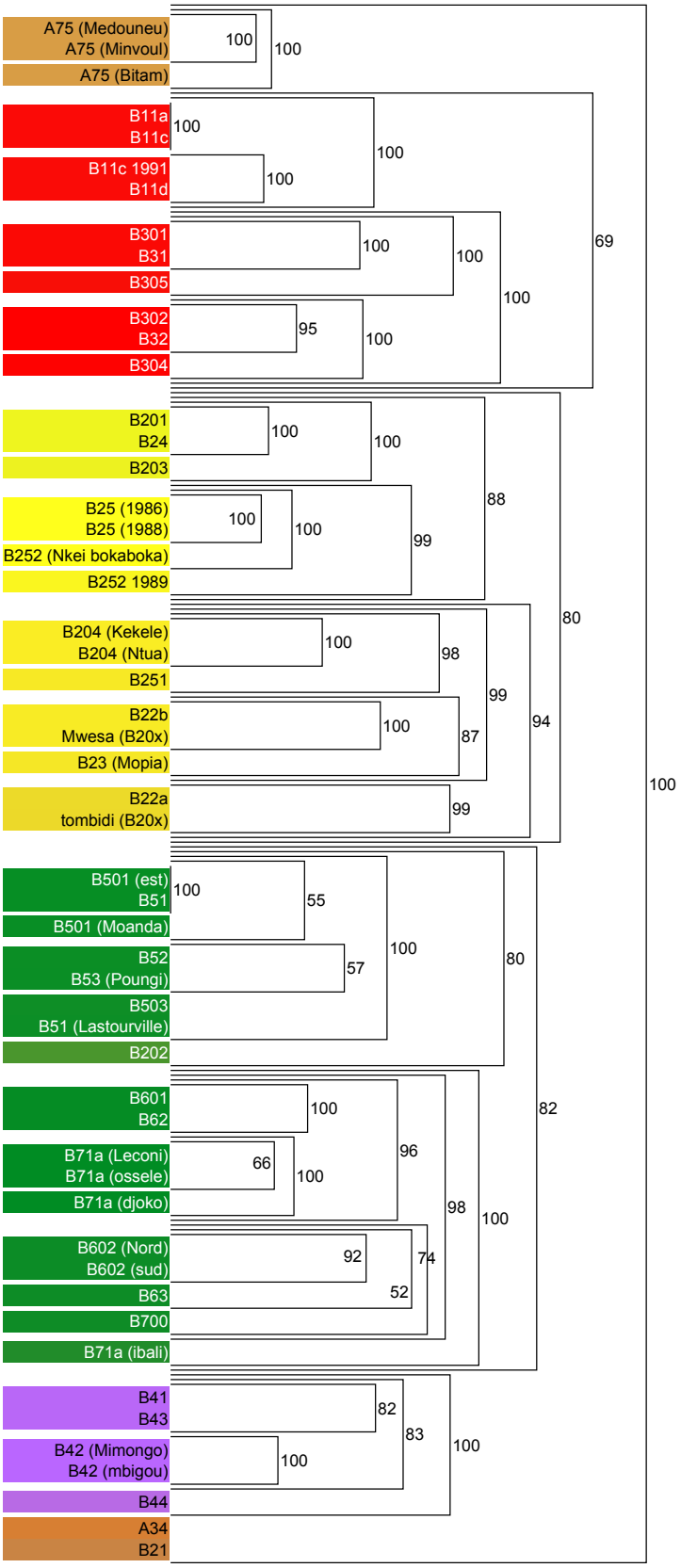


(a) Phonetic distances for all location pairs; dark is close, light is far.

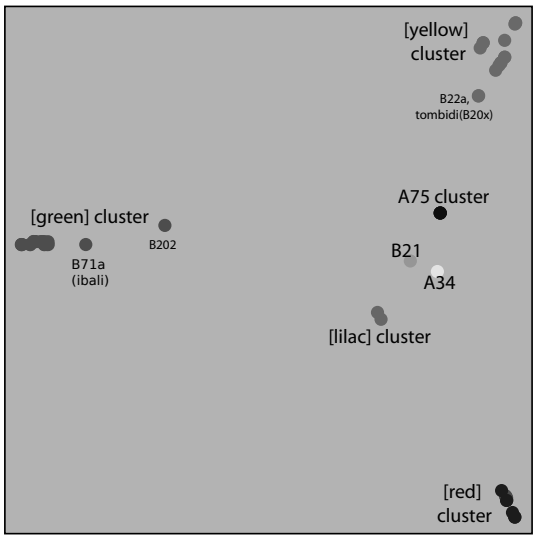


(b) This map displays the first three dimensions of the results of MDS applied to the mean cophenetic distances used in the bootstrapped clustering procedure.

Figure 3: Result maps



(a) Consensus dendrogram. For reference, the labels are colored as in Fig. 3(b)



(b) MDS plot: a scatterplot that shows the result of two-dimensional MDS on the bootstrapped data. A few relative outliers are marked; compare with the consensus diagram

Figure 4: Result diagrams

Appendix: Guthrie codes - language code reference

For reference, a list of which Ethnologue language codes[9] the Guthrie codes correspond to:

Guthrie code	Name	Ethnologue/ISO code
A34	Benga	[bng]
A75	Fang	[fan]
B11a	Mpongwe	[mye]
B11c	Galwa	[mye]
B11d	Dyumba	[mye]
B201	Ndasa	[nda]
B202	Sigu	[sxe]
B203	Samay	no code available (not listed as such)
B204	Ndambomo	no code available (not listed as such)
B21	Seki	[syi]
B22a	Kele	[keb]
B22b	Ngom	[nra]
B23	Mbangwe	[zmn]
B24	Wumbvu	[wum]
B25	Kota	[koq]
B251	Shake	[sak]
B252	Mahongwe	[mhb]
B20x	Mwesa	no code available (not listed as such)
B20x	Tombidi	no code available (not listed as such)
B301	Viya	no code available (not listed as such)
B302	Himba	[sbw]
B304	Pindji	[pic]
B305	Vove	[buw]
B31	Tsogo	[tsv]
B32	Kande	[kbs]
B41	Sira	[swj]
B42	Sangu	[snq]
B43	Punu	[puu]
B44	Lumbu	[lup]
B501	Wanzi	[wdd]
B503	Vili	no code available (not listed as such)
B51	Duma	[dma]
B52	Nzebi	[nzb]
B53	Tsaangi	[tsa]
B602	Kaningi	[kzo]
B62	Mbaama	[mbm]
B63	Ndumu	[nmd]
B700	Tsitsege	[tck]
B71a	Tege	[teg]

[9] See also ISO 639-3, though Ethnologue updates its reference more often