Introducing a state-of-the-art phylogenetic classification of the Kikongo Language Cluster

Gilles-Maurice de Schryver, Rebecca Grollemund, Simon Branford and Koen Bostoen

Abstract

It is shown how past lexicostatistic efforts eventually led to lexically-driven phylogenetic classifications of the Bantu languages. As a new case study, 95 North-West and West Bantu language varieties are sampled across geographical space, with a focus on the wider Lower Congo region. This leads to the discovery of a discrete clade within West-Coastal Bantu, which we term the Kikongo Language Cluster (KLC), a disparate continuum of closely related Bantu languages. Both a branching tree and a continuum model are called in to ‘define’ the true nature of the KLC, and pre-historical implications are drawn from this.

Keywords: Kikongo, Bantu, lexicostatistics, phylogenetic classification, Bayesian MCMC, Neighbour-Net
Introduction: Aim, methodology and value of the study

The aim of the present study is twofold. On the one hand we wish to present the first character-driven phylogenetic classification applied to lexical data of the Kikongo Language Cluster (KLC), and on the other we wish to present an exhaustive overview of the field of Bantu lexicostatistics as far as it pertains to the KLC. Indeed, our novel classification—which not only pays close attention to the interactions of the KLC with its linguistic neighbours through an external classification but also, and especially, to the divisions within the KLC itself through an internal classification—follows in the footsteps of an illustrious tradition of nearly half a century of lexicostatistic classifications of the Bantu languages undertaken by various linguists at Tervuren and elsewhere. The major output from Tervuren is the set of distance-based language trees and maps for Bantu found in Bastin et al. (1999), built up from lexical similarity measures across the basic vocabulary of up to 542 present-day Bantu (and Bantoid) language varieties. Our approach also starts from basic vocabulary, and for reasons of interoperability, even the same basic-vocabulary items as in the Tervuren studies. But rather than building trees based on lexical similarities—which produce mere distance matrices, and implicitly assume a constant rate of lexical change—as is customary in lexicostatistics, our advanced computational models focus on the study and evolution of the actual characters. We will also strive to employ a method which represents phylogenetic uncertainty in the resulting tree itself. This will enable us to stop pretending that we know the exact phylogeny of any Bantu tree, but allows us to instead state the truth that what we propose is the best possible phylogenetic hypothesis that can be made with today’s data sets and computational tools. Another important methodological point is that we will use synchronic lexical data only (up to maximum half a century old). From a Bantu-wide perspective this may look like an obvious statement, but specifically for Kikongo it is not, given the availability of language data in a few historical Kikongo varieties going back a number of centuries (Bostoen & de Schryver 2015). Bringing such diachronic data into our models will be pursued in a follow-up article.

In the previous paragraph we have referred to the Kikongo ‘Language Cluster’, but apart from uses in this introduction, we will only do so again once such a cluster has become apparent from our phylogenetic classification. At the outset, all we assume to know is that we are dealing with a number of Kikongo varieties spoken in the Lower Congo region, on both sides of the mouth of the Congo River, in four present-day countries: Angola (in the north as well as in Cabinda), the Democratic Republic of the Congo (in the Lower Congo province), the Republic of the Congo or Congo-Brazzaville (in the south), and Gabon (along the coast). For now, we may say that we are dealing with Guthrie’s H10 group (cf. Section 1.2.1, Figure 8), but part of the research is to confirm whether all H10 varieties indeed belong together, and to determine whether, and if so which, other varieties (in other H groups or even other zones) ought to be seen as Kikongo(id). Stated bluntly, we wish to find out how large the KLC is. As a starting point, however, we may zoom in on the central region as done in Figure 1: Bantu varieties spoken in the Lower Congo region., and ask ourselves which of the shown Bantu varieties (in orange)—both those mapped
and those in the text balloons—actually are Kikongo(id). Note that we are not going to concern ourselves with today’s vehicular Kikongo, known as Kikongo ya Leta in the DRC, or Munukutuba in Congo-Brazzaville (Samarin 2013). Nor do we want to simply equate Kikongo with the area of the former Kongo kingdom (Pigafetta 1591, Cavazzi da Montecuccolo 1687, Jordão 1877, Cuvelier 1941, 1946, Balandier 1965, Randles 1968, Bontinck 1972, Ekholm 1972, Thornton 1983, Hilton 1985, Thornton 2001). There are certainly Kongo people who identify themselves with Kikongo on cultural and/or ethnic grounds, but what interests us here is the linguistic evidence.

Language classificatory tasks are rarely produced in isolation, as Guthrie pointed out: “nous nous attachons aux problèmes de classification essentiellement afin d’en utiliser les résultats pour d’autres études” (Guthrie 1959:73). The classifications of the Bantu languages—lexicostatistic and other ones—have mainly been undertaken with the aim to obtain insights into the history of the people currently speaking these languages. In this respect, there is a great divide between those genealogical classifications in which there is a primary split between East and West Bantu, vs. those where the deepest splits on the tree are within West Bantu, namely in the north-western zones A and B, with East Bantu branching from West Bantu at a later stage. This divide has resulted in two very different models of evolution (Wiesmüller 1997, Bostoen 2004, Bostoen & Grégoire 2007). In the ‘East parallel to West’ model, East Bantu is a primary branch of Proto-Bantu and its arrival in eastern Africa is the result of an eastward dispersion from Cameroon along the northern borders of the rainforest, towards the Great Lakes region. West Bantu is the result of a north-south movement through the equatorial rainforest and a
rapid disintegration into different sub-groups. In contrast, in the ‘East out of West’ model, East Bantu is only a relatively late Bantu offshoot and constitutes a greater sub-group known as ‘Savannah Bantu’ together with, roughly, the West-Coastal Bantu and South-West Bantu sub-groups, as opposed to the remainder of the Bantu languages, the so-called ‘Forest Bantu’ languages. An early proponent of the primary split between East and West was Guthrie (1967-71), and it may also be found in Bastin et al. (1983). This view is not supported by for instance the classifications of Heine (Heine 1973, Heine et al. 1977) nor even Henrici (1973) who reprocessed Guthrie’s own data. Knowing which one it is is important, as it leads to very different historical accounts: In Vansina (1984, 1990) the point of departure is the primary split, for Ehret (1998) the deepest splits are within West Bantu. The ‘East out of West’ model is also seen in the latest work from Tervuren (Bastin et al. 1999), and it may be seen in most phylogenetic studies (Holden 2002, Holden et al. 2005, Pagel & Meade 2006, Rexová et al. 2006). Occasionally, however, the ‘East parallel to West’ model, or a variation thereof, reappears in phylogenetic studies (e.g., Holden & Gray 2006:21). Linked to these two models is the so-called ‘Kongo-Nukleus’ (Heine 1973, Heine et al. 1977), whereby the Lower Congo region is seen as a secondary nucleus of Bantu expansion. For instance, in Heine et al. (1977) the Bantu languages are divided into eight coordinate groups, with the eighth called ‘Kongo branch’. The third sub-group within that branch is the ‘Kikongo group’, while the last sub-group within the Kikongo group is further subdivided into another 26 sub-groups, containing about 200 languages, mainly from Guthrie’s eastern Bantu. Understanding the external classification of the greater Lower Congo region is thus of paramount importance.

Obtaining a better grip on the internal classification of the Lower Congo region is as important. Barring one study limited to the DRC (Ntunda Nzeza 2007), no work has been published on the internal classification of the KLC. Instead, a single or at least a limited number of Kikongo varieties were chosen rather randomly in various classification studies, and ‘Kikongo’ was treated as ‘one language’, typically ‘H16’. Among the claims one finds in the literature (e.g., Obenga 1970, Pinçon & Ngoïe-Ngalla 1990, Futi 2012:19) is that the wide distribution of Kikongo varieties is the result of the spread of the Kongo kingdom. The many Kikongo ‘dialects’ (Laman 1928, 1936) would then be the result of the collapse of that kingdom in the 17th century. Some scholars even date the fragmentation to just over a century ago. According to Ntunda Nzeza (2007) all the Kikongo varieties from the DRC were a single unitary language until very recently indeed, with a split between central and peripheral Kikongo varieties appearing near the end of the 19th century only, when the centre (in the DRC) would start to ‘push back’ any new cultural traits coming from the peripheries:

Les voies de communication créées par les explorations de Stanley en 1877 et 1879, par le système de portage (institué vers les années 1885) entre Léopoldville « devenu un centre de dépôt », et la côte atlantique, par la construction du chemin de fer entre Matadi et Léopoldville (de 1890 jusqu’à 1898) ainsi que par les pistes des caravanes que fréquentaient des personnes d’origines diverses..., auraient favorisé le refoulement de plusieurs traits culturels du centre vers les Zones périphériques de l’espace koongo.
(Ntunda Nzeza 2007:106)
The present study is part of a series of studies with which we hope to start to provide an answer to the question as to whether today’s Kikongo varieties are indeed that young (i.e., originated at the end of the 17th century, or even at the end of the 19th century), or whether the break-up dates from much earlier.

The main linguistic exercise being the production of a new language tree for Kikongo, it should be emphasised that the present research is not a dialect(ometrical) study: no phonological or other distances will be calculated and no dialect maps will be drawn up (so no approaches, methods nor references from that field are included either). Rather, we will include many (including ‘new’) Kikongo varieties in our study, together with all the immediate neighbours and the neighbours of those neighbours. The chosen varieties are representative of the geographical diversity; this research is thus a comprehensive approach to the entire KLC. In Section 1 an overview is given of the existing classifications for Bantu and Kikongo, with sub-sections for the historical, lexicostatistic and phylogenetic classifications. Given that this section serves as a detailed grounding for our own research, various techniques are also discussed in the process. Our own research takes up the remainder of this article: with materials and methods in Section 2, results in Section 3, a discussion in Section 4 and finally, a conclusion.

1. Classifying Bantu and Kikongo: State of the art

1.1. Past classifications of the Bantu languages

Excellent overviews of the various past classifications of the Bantu languages can be found in Doke (1943), Bastin (1978) and Heine (1980). More recent overviews, which cover the research highlights of the past few decades, include Hinnebusch (1989), Nurse (1994-95), Nurse & Philippson (2003), Marten (2006) and Grollemund (2012). Grollemund (2012) also summarises the history of the place of (Narrow) Bantu within the Niger-Congo family, and within that, the Bantoid (i.e., Wide Bantu) language family. The interested reader is referred to these works, as only those classifications with a direct bearing on the present study will be discussed in the following.

1.2. Past classifications of the Kikongo varieties

1.2.1. Presentation

Remarkably, fifty years before the very first internal classifications of (what later became called) the Bantu language family, and nearly a century before Bleek (1857) coined the term ‘Bantu’ itself, Proyart (1776) had this to say about the languages spoken in the various kingdoms around the mouth of the Congo River:

L’idiome de Kakongo, le même à-peu-près que celui de Loango, n’Goio, Iomba & autres petits Etats circonvoisins, diffère essentiellement de celui de Congo. Plusieurs articles semblables, & un grand nombre de racines communes, semblent cependant indiquer que ces langues ont eu la même origine; mais on ignore laquelle des deux est la langue mère. Les plus habiles d’entre les Negres n’ont pas la moindre idée de l’origine ni des progrès de leur langue: ils parlent, disent-ils, comme ils ont oui-parler leurs peres. (Proyart 1776: 171-2)
Here, Proyart is in effect not only hinting at a common parent for a number of Bantu languages, but already making a first internal division within the Kikongo group, i.e., those Kikongo varieties spoken north of the Congo River vs. the one spoken to the south of it. Figure 2 reproduces Proyart’s accompanying map.

**Figure 2.** The kingdoms around the mouth of the Congo River during the second half of the 18th century (Proyart 1776).

In the first two Bantu-wide classifications, both published in 1826, the so-called Kongo languages stand out as a major group. In Prichard’s (1826) classification, the ‘Kongo Race’ is one of just three groups, the other two being the ‘Kaffers’ and the ‘Mosambique Nations’. Within the ‘Kongo Race’, a first sub-group includes ‘Kongoes, Angolans and Dongo People’, together the ‘Kongo Language’, a second the ‘Loango People, Makongo and N’Goy’, together the ‘Loango Dialect of the Kongo’. The Congo River again divides the two sub-groups. In Balbi’s (1826) classification of the languages of Africa, the ‘Congo Family’ shows a close affinity indeed, bringing together Loango, Camba, Anzico (or Makokko), Congo, Bunda (or Angola), Benguela, Mandongo and Molua. If the series of articles published in 1847 by Appleyard (1847a, 1847b, 1847c, 1847d, 1847e) is to be seen as a classification, then his divides the Bantu languages into five groups: ‘Congo’, ‘Damara’, ‘Sechuana’, ‘Kaffir’, and ‘Unclassified’. Under ‘Congo’ he only recognizes Congo, Bunda and Benguela.
Of the eleven major African-language groups (plus a supplementary one for all the unclassified and isolated varieties) in Koelle’s (1854) inventory, Group X contains the ‘Kongo Ngola’ languages. Within Group X there are three sub-groups (A to C), with respectively seven, five and six members. Five of these can indeed be shown (cf. Bostoen & de Schryver 2015) to be Kikongo varieties: X.A.1 Kabênda (i.e., Kakongo), X.A.2 Mimbôma (i.e., Kimboma), X.A.3 Musentându (i.e., Kintandu), X.B.4 Nyömbe (i.e., Kiyombe) and X.B.5 Basúnde (i.e., Central Kisundi). Four of these five are also placed rather accurately on the accompanying map for the whole of Africa, as may be seen from the extract in Figure 3.

**Figure 3.** The geographical location of four of the five Kikongo varieties for which Koelle (1854) collected data.

The last major classification of the 19th century may be found in Bleek’s (1862, 1869) comparative grammar of the South African languages, where his ‘Kongo genus’ consists of the southern Kongo and the northern Kakongo, but also includes Mpongwe, spoken on both sides of the Gabon estuary way to the north.

In Johnston’s (1922) classification of the Bantu and semi-Bantu languages, Group Z is defined as follows:

**GROUP Z. KOŃGω**

(100) Kisi-końgω [São Salvador and Boma]
   (100a) Ki-sorońgω or Ki-solońgω or Ki-sonyω
   (100b) Ki-końgω [including the ‘dialects’ of the Ba-bwendi, Ba-sundi and Ba-sese]
(101) Kakońgω [spoken in Kabinda]
(102) Ki-yombe or Ci-luańgω
   (102a) Ki-vili or Ki-vumbu or Lu-wumbu (N. Luańgu)
(103) Ki-lumbo of Mayumba
(Johnston 1922:7, 105-6; Note that (103) is actually a dialect of (102a), not (102a) of (102) — cf. errata on p. x)
The section of the map illustrating the geographical spread of these varieties is reproduced in Figure 4, with Johnston (1922:105) pointing out that “The Lower Congo river is almost a line of demarkation between them [i.e., between (100) vs. (101) to (103)].”

**Figure 4.** The geographical location of Johnston’s (1922) ‘Koŋgo group’, i.e., the southern (100), (100a) and (100b) vs. the northern (101), (102), (102a) and (103).

The first detailed dialect map of the various Kikongo varieties appeared in Laman’s (1936) Kikongo-French dictionary. While this dictionary mainly covers the variety Laman is thoroughly acquainted with—a variety which he saw as and called the ‘central’ one—a painstaking perusal of all earlier Kikongo dictionaries enabled him to add to his own dictionary what he then called the ‘dialectal forms’. Throughout his dictionary, Laman uses the following codes, which are also depicted on the accompanying map shown in Figure 5.

- **C** central (around Mukimbungu)
- **S/SB** south (down to San Salvador)
- **E/EB** east (around Kisantu)
- **SE** south-east
- **NE** north-east (north of Brazzaville)
- **N** north (around Kingoyi)
- **Be** Bembe dialect (around Mouyonzi)
- **NO** north-west, Kunyi dialect (around Loudima)
- **Vi** Vili dialect (around Loango)
- **Ka** Kakongo (south of Loango)
• Nd  Ndingi or Ndinzi (north-western Mayombe, on the border with Cabinda)
• Mb  Mboka (spoken in Cabinda)
• O  west, Yombe dialect (My)

Figure 5. Laman’s (1936) dialect map of the Kikongo varieties.

In his dictionary, Laman (1936:xl-xci) gives the phonetic characteristics of each dialect, but he stops short from sub-grouping those dialects in any way.

The most comprehensive enumeration of Kikongo languages and dialects may be found in van Bulck (1948:328-92), who lists about a hundred historic and then current varieties. A selection of those was included on his accompanying map, and is reproduced in Figure 6. That selection corresponds well with his ‘Kɔɔŋgɔ group’ in van Bulck (1954), where the number of varieties was brought down to only contain about fifty, as shown below. Wherever possible, the 1948 map numbers were added to the 1954 enumeration—in bold, following each variety. The numbers of the Kikongo varieties that were not mapped are between brackets; these include varieties spoken outside the then Belgian Congo as well as a large number of the then current ‘dialects’ (their numbers start with ‘172/...’). For two varieties no number was found.
**Figure 6.** Excerpt from van Bulck’s (1948) language map of the then Belgian Congo; with a focus on his ‘Kɔɔŋgɔ group’ (in green).

**Kɔɔŋgɔ group**

1. Kwango sub-group
   a) kiZɔɔmbɔ
      a. Tumba Mani dialect: kiZɔɔmbɔ 167b
      b. Nsɔɔ dialect: kiNsɔɔ (173)
   b) kiSuku
      a. Angolan dialect: left bank of the Kwango 142a
      b. Lukula dialect: Gingungi 142b
      c. speech of the baDɔɔndi (no.?)
      d. Tungila dialect: South kiSuku 142c
   c) kiTsam: kiTsamba dialect 236
      kiHuŋgaan: kiHuŋgana dialect 237

2. Mouth of the Congo River sub-group
   a) kishi(ma)Luangu dialect 174
      kishiKakɔɔŋgɔ dialect 159
      kishiKabinda dialect (172/25)
      Ndingi dialect (172/26)
      Mboka dialect (172/27)
   b) kisi(ma)Yɔɔmbε dialect 149
      kiVungunya dialect 157
      ‘classic’ kiYombe (172/30)
   c) kisɔLɔɔŋgɔ dialect 144
      kiVili = kiPidi 145
      kiWɔɔɔɔ dialect 146
   d) kiSuundi (northern speech: Dganda
      Tsuundi 147a; western speech 147b;
      eastern speech: kisimaNyanga 148;
      southern speech: Ngombe-Matari 147d)
      kiDɔɔndɔ (150)
      kiKaamba 151
      kiBɛɛmbɛ 152
      kiInda dialect 153
      kiGaangala 154

3. Inkisi sub-group
   a) Eastern dialects
      kiKɔɔŋgɔ: dialect of the
      baMpaangu (kiNtaandu speech 161b;)
      kiMbaamba speech (172/11); Kimpese
      speech 163a; Luula-Lusanga speech
      kiPiindi 238 (on the Nsele) 170a
   b) South-East dialects
      kiKɔɔŋgɔ of Angola
      (North-East) (172/17)
      kiʃiKanu 168a
      kiMbeble (Mbeko-North speech
      (Mbeko-tsɛkɛ) 168c; Mbeko-South
      speech 168b)
      kiPhatu 169

4. Lower River sub-group
   a) Southern dialects
      kifiKɔɔŋgɔ of San Salvador (172/14)
kiMbata dialect: Mbata Makela 167a
Kinzaamba speech (172/15)

b) Central dialects
kiKɔɔŋgɔ: Matadi (Mazinga speech)
(Mpalabala); Mukimvika speech) 172b
kiKɔɔŋgɔ-kiMboma: Mboma 155
‘unified’ Mazinga-Bɛɛmbɛ (172/5)
kiNdibu (left bank of the Inkisi) 166

5. French Equatorial Africa sub-group
kiYaka of Kwilu-Niari 209
(van Bulck 1954:263-5)

In 1950 Hulstaert published his language map of the then Belgian Congo, together with an accompanying booklet of 67 pages. The map itself was produced in 1948 and is dated 1948. Incongruously, rather than referring to languages, Hulstaert refers to people (Hulstaert 1950:14), so his labels are typically in class 2 rather than in class 7. His ‘group 23’, shown in Figure 7, are the Bakɔɔŋgo.

Figure 7. Excerpt from Hulstaert’s language map (produced in 1948) of the then Belgian Congo; with a focus on his ‘Bakɔɔŋgɔ group’ (in green, numbered 23).

1. An interesting comparison of Hulstaert’s map with van Bulck’s of the same year may be found in Van de Velde (1998), which was summarised and translated into English in Van de Velde (1999).
The Bakɔngɔ include the (a) Bayɔmbɛ, (b) Bambala [in the Kwilu-Kwango], (c) Bayaka, (d) Masɔndɛ, (e) Bangongo, (f) Balua, (g) Basuku, (h) Batsaam, (i) Bazombo, (j) Bampangu, (k) Bambata, (l) Manyanga, (m) Basɔlɔngɔ, (n) Baholo, (o) Basuku-Sud, (p) Bakwɛsɛ, (q) Tupɛndɛ, and (r) Bambundu. Observe thus that the Kikongo region was extended all the way to Loanda “où il semble qu’on doive y rattacher les Bambundu” (Hulstaert 1950:33). Also shown on the map within this group are the Bakunyi, Babeembe, and Balari, but these are not listed in the map’s legend under group 23 nor numbered. Not shown on the map, but mentioned in the accompanying booklet as belonging to group 23 are also ‘the dialects of Cabinda’ and ‘the Civili of Loango’ (Hulstaert 1950:34). Further note that Hulstaert also extended his Bakɔngɔ region beyond the north of Pointe Noire, along the Atlantic coast.

In Guthrie’s (1948) classification of the Bantu languages—a mostly geographic-referential system, underpinned by some typological features—the Kikongo varieties are found within zone H, more specifically group 10 within this zone, as shown below as well as on the extract from his accompanying map in Figure 8.

---

Figure 8. Guthrie’s (1948) zone H on his geographic-referential map of the Bantu languages.²

---

2. Interestingly, prior to the publication of Guthrie’s (1948) classification, Doke (1945), in the context of a bibliographical survey, also classified the Bantu languages using zones and groups, as well as clusters (i.e., languages) and dialects, in that order (see also Cole 1951, 1959, Cope 1971). With regard to Kikongo, Doke (1945:16-22) refers to the usual north-south divide as well as Laman’s (1936) detailed classification.
ZONe H

GROUP 10
H.11 Vili, ki-
H.12 Kunyi
H.13 Bembe
H.14 Ndingi
H.15 Mboka

GROUP 20
H.41 Mbala, ki-

GROUP 30
H.16 KODGO
H.16a E. KODGO, ka- (Fiote)
H.16b YOMBE, ki-
H.16c SUNDI
H.16d BWENDE
H.16e N.E. KODGO, ki-
H.16f KODGO, ki-
H.16g S. KODGO, kisi-
H.16h ZOMBO, ki-

GROUP 40
H.42 HUDANA, ki- (Huana)

(Guthrie 1948:79)

In all fairness it must be pointed out that Guthrie was not too sure about the lumping of the typological features of the languages from groups H20, H30 and H40 on the one hand with those from H10 on the other into one zone H:

The difficulties of this zone are great, in view of the peculiar one-sidedness of the available data. On the one hand the KODGO (16) dialects are well known, as to both vocabulary and structure, but apart from NDODGO (21) we are faced on the other hand with an almost complete lack of information. This means that the grouping is unreliable, and also that any description of the zone must be scrappy. From what we do know about these languages it seems reasonable to put these four groups into one zone [...]

(Guthrie 1948:51)

Following a fieldtrip in 1949, Guthrie was able to revisit his classification for the zone A and B languages, which resulted in his (1953) reclassification of the Bantu languages of western equatorial Africa. The southernmost of these being adjacent to H10, with especially his B40 Shira-Punu group important for our study, this reclassification is presented here, together with the accompanying map excerpt in Figure 9.

B.40 SHIRA-PUNU
B.41 SIRA, i- [Shira]
B.42 SANGU, yi- [Shango]
B.43 PUNU, yi-
B.44 LUMBU, i-

(Guthrie 1953:76)
Figure 9. Excerpt from Guthrie’s (1953) reclassification of the Bantu languages of western equatorial Africa; with a focus on the ‘B40 Shira-Punu group’.

Gossweiler’s (1953) map of the languages of Angola includes ‘Kikongo’, see Figure 10, for which the following varieties are listed: Maiombe, Kakongo, Cabinda, Mussurongo, Muxikongo, Muzombo, Kikongo, Sosso, Mussuco.

Figure 10. Excerpt from Gossweiler’s (1953) map of the languages of Angola; with ‘Kikongo’ numbered ①.

Compiling all the ethnolinguistic data available up to that point in time, Boone (1973) produced an ethnographic map of the then Zaïre. Figure 11 shows the Lower Congo section.
Figure 11. Boone’s (1973) ethnographic map of the then Zaïre, Lower Congo section.

Over the years, Guthrie kept changing around his language varieties and codes, lumping, adding and deleting material, so that by the time his four-volume *Comparative Bantu* appeared (Guthrie 1967-71)—a mere two decades after his first classifications (Guthrie 1948, 1953)—his initial reference system had all but broken down for quite a number of groups. This is also the case for Guthrie’s H10 (regarding the northern H10 varieties, compare for instance Figure 8 with Figure 9), for which the new pairings of codes and language varieties are shown in the first two columns of Table 1. As a consequence, the ‘codes as geo-location’ of especially Guthrie’s initial (1948) map, cf. extract in Figure 8, has become almost unusable. A few years later Bastin (1978) revised Guthrie’s referential classification yet again, with the changes and extensions for H10 as shown in the next two columns of Table 1. The most recent update of Guthrie’s classification is Maho’s (2009), illustrated for H10 in the last two columns of Table 1. Also shown in Table 1 are the H10 varieties selected by Bastin et al. (1999) for their Bantu-wide classifications (cf. Section 1.3.2 below).
Table 1. A comparison of the referential classifications for H10 of Guthrie (1967-71), Bastin (1978) and Maho (2009), together with the H10 languages selected by Bastin et al. (1999). [The data is aligned based on the codes; a dash ‘-’ means ‘no correspondence’.]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Code Variety</td>
<td>Code Variety</td>
<td>Code Variety</td>
<td></td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>H11 Bembe</td>
<td>H11 Bembe (= Beembe)</td>
<td>H10 Mboma</td>
<td>H11 Bembe, including Keenge</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>H11 Bembe</td>
<td>- -</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>H11 Hangala, Ghaangala</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>H11 Bembe, including Keenge</td>
<td></td>
</tr>
<tr>
<td>H112 Vili</td>
<td>H112 Vili (= Ci-vili = Fiote)</td>
<td>H12a Vili</td>
<td>H12 Vili, Civili</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>H12a Vili</td>
<td>- -</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>H12b Yombe († West Kongo)</td>
<td>H12b Yombe</td>
</tr>
<tr>
<td>(Vunginyi = Vungunya, Suundi)</td>
<td>- -</td>
<td>H12b Yombe</td>
<td>- -</td>
</tr>
<tr>
<td>H13 Kunyi</td>
<td>H13 Kunyi</td>
<td>- -</td>
<td>H13 Kunyi</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>H13a Kunyi</td>
<td>- -</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>H13b Suundi</td>
<td>- -</td>
</tr>
<tr>
<td>- -</td>
<td>- -</td>
<td>H13c Nyanga</td>
<td>- -</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------</td>
<td>----------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>H14 Ndingi</td>
<td>H14 Ndingi (= Ndinzi = Ngingi)</td>
<td>-</td>
<td>H14 Ndingi (Cabinda), Ndinzi</td>
</tr>
<tr>
<td>H15 Mboka</td>
<td>H15 Mboka</td>
<td>-</td>
<td>H15 Mboka (Cabinda)</td>
</tr>
<tr>
<td>H16 Kongo cluster</td>
<td>H16 Kongo</td>
<td>-</td>
<td>H16 Kikongo, Kongo</td>
</tr>
<tr>
<td>H16a South Kongo</td>
<td>H16a South Kongo (Mbata, Nzamba)</td>
<td>-</td>
<td>H16a South Kongo, including Kisikongo, Mboma, Songo, Solongo</td>
</tr>
<tr>
<td>H16b Central Kongo</td>
<td>H16b Manyanga (= Central Kongo) (Mboma, Ngoyo, Ndibu, Solongo, Kongo of Mpalabal, Kongo of Mazingo, Kongo of Mukimvika)</td>
<td>H16b Manyanga</td>
<td>H16b Central Kongo, including Suundi, Mazinga, Manyanga</td>
</tr>
<tr>
<td>H16c Yombe</td>
<td>H16c [Yombe &gt; cf. H12b]</td>
<td>-</td>
<td>H16c Yombe</td>
</tr>
<tr>
<td>H16d West Kongo</td>
<td>H16d West Kongo (= Fiote)</td>
<td>-</td>
<td>H16d West Kongo, including Woyo, Fiote, Kakongo, Kako</td>
</tr>
<tr>
<td>H16e Bwende</td>
<td>H16e Bwende (= Ngoyo = Fiote = Buende)</td>
<td>-</td>
<td>H16e Bwende, including Sonde</td>
</tr>
<tr>
<td>H16f Laadi</td>
<td>H16f Laadi (= Lari), Mbinsa</td>
<td>H16f Laadi</td>
<td>H16f Laadi</td>
</tr>
<tr>
<td>H16g East Kongo</td>
<td>H16g Ntandu (= Central-East Kongo = Phaku = Mphaku) (Mpaangu, Mbaamba, Mpese, Luula)</td>
<td>H16g Ntandu</td>
<td>H16g East Kongo, including Santu, Ntandu</td>
</tr>
<tr>
<td>H16h South-East Kongo</td>
<td>H16h Kongo of San Salvador (= South Kongo = Xikongo = South-West Kongo)</td>
<td>H16h Sikongo</td>
<td>H16h South-East Kongo, including Nkanu, Zoombo, Pende</td>
</tr>
<tr>
<td>Code</td>
<td>Variety</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H16i</td>
<td>Sundi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H16i</td>
<td>Sondi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H16j</td>
<td>Hangala</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H16j</td>
<td>Kamba</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H16k</td>
<td>Zombo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H16j</td>
<td>Zombo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H16k</td>
<td>Kamba</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H17</td>
<td>Dondo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H17a</td>
<td>Dondo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H17k</td>
<td>Kaamba</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
With a focus on the DRC, the most recent inventory for the Kikongo varieties is included in Kadima et al. (2008), who recognize eight main varieties together with their dialects, as follows:

[601] kiwoyo (cikongo, kongo); [602] kisolongo (asolongo), kikongo des Bakongo ya Boma; [603] kiyoombe (yoombe, yombe) - de Kakongo - de Kwimba ou Tsanga nord & sud - de Madúda - de Phatu - de Tsundi - de Vaku - de Singini - de Tshela - de Seke-Banza; [604] kimboma - de Kionzo - de Mbanza-Manteke (parler des ressortissants du plateau Mbanza-Manteke) - de Nduizi; [605] kimanyanga (manyanga, manyaanga) - de la kenge - de Kimbanza - de Kimumba - de Kinkenge - de Kivunda - des Balari - de Mbanza-Mona - de Mbanza-Mwembe - de Mbanza-Ngoyo - de Mongo-Luala - de Kasi - de Isangila; [606] kindibu (ndibu) - de Mbanza-Ngungu et Nkolo - de Kwilu-Ngongo - de Kimpese - de Nkolo - de Songololo (voisin de Kisansala de l’Angola) - de Bangu (Besingóómba) - de Kimpangu (voisin de Kizombo de l’Angola); [607] kisingóómba (anciens bansundi, manyanga ma Loonde) - kimayaamba - kinsuündi (voisin du kilémfu); [608] kintándu (ntandu) - kilémfu - kimbáta (voisin du kizómbo de l’Angola) - kimbéku - kimpángu (voisin du kizómbo de l’Angola) - kinkánua (début du kiyáka) - kimpése - kintándu (dialecte de Mbetengi) - kinsuündi (début du kilémfu)

(Kadima et al. 2008:14-5)

On their language maps (Kadima et al. 2008:143, 184), an average of two points per main variety are shown.

1.2.2. Discussion

Given the Kongo region’s early contact with the western world, it is not surprising to see its languages featured in and from the very first classifications for the Bantu languages onwards. At the same time, in terms of internal classification, one cannot but fail to conclude that not that much progress was made since Proyart’s (1776) early observation that the Congo River divides the northern varieties spoken in the then Kakongo, Loango, Ngoyo, Mayombe and other kingdoms north of the river, from the southern variety spoken south of the river in the then Kongo kingdom. Certainly, numerous ‘dialects’ have been added to the inventory over the years, but overlooking a list like van Bulck’s (1954), one suspects an all too easy equation of ‘village-lect’ with ‘dialect’. While the first scientific attempt to map the Kikongo dialects is found in Laman’s (1936) dictionary, also that effort was largely opportunistic, as driven by the other dictionaries that happened to have been compiled by then: Bentley (1887, 1895) for the southern variety around Mbanza Kongo, Anon. (1890) for the Kakongo variety in Cabinda, Butaye (1909) for the eastern variety around Kisantu, Marichelle (1912) for Civili, and Bittremieux (1922, 1927) for Kiyombe. Throughout his dictionary, material labelled as SB literally stands for the South ‘Bentley’ dialect, and EB for the East ‘Butaye’ dialect. Laman’s synthesis was useful to Guthrie, though, cf. Guthrie (1948:5).

Given all the earlier scholarship, there are a few striking absences from Guthrie’s H10, however. Take for instance Kisolongo, already mentioned by Johnston (1922),
where Kisolongo (100a) is seen as a dialect of Kisikongo (100). Kisolongo was not included on Laman’s (1936) map either (there were no dictionaries of Kisolongo to draw from), so it may not be surprising that it is also absent from the classifications of Guthrie. Kisolongo is ‘re-included’ by van Bulck (1954), in his sub-group 2(c), where he joins it to the western Civili and Ciwoyo however. In Bastin (1978) Kisolongo is seen as a central variety, found in H16b, together with varieties such as Kimanyanga and Kindibu. In Maho (2009), finally, Kisolongo re-joins the southern group with Kisikongo, i.e., H16a—nearly a century after Johnston (1922). This exercise may be repeated for a number of other Kikongo varieties, but the point is clear: there is a dire need to arrive at a scientifically sound internal classification of the Kikongo varieties. While Guthrie’s codes may continue to be used—through their latest Maho (2009) incarnation—the codes are just that: geo-location shorthands to refer to certain Bantu language varieties, with no classificatory value whatsoever.

1.3. Lexicostatistics of the Bantu languages

1.3.1. Overview

The linguistic classifications referred to in Section 1.1 and those presented in Section 1.2 were all based on geographic proximity and/or the distribution of certain phonological, grammatical or other linguistic characteristics. The study of such distributions may lead to the reconstruction of parent and even proto forms, with which one attempts to build a genetically valid tree. This comparative method has so far failed to produce a satisfactory classification for the entire Bantu language family, however, presumably “due to the complexity of the task involved” (Marten 2006:46). The comparative method has nonetheless resulted in convincing sub-classifications of smaller sub-groups, such as Nurse & Hinnebusch’s (1993) sub-classification of the Sabaki languages. In the words of Marten (2006:46): “what needs to be done is to work on smaller subgroups, and once these smaller groupings are better understood, construct a larger tree upwards, until all Bantu languages are included”. That said, not everyone believes that it is even theoretically possible to build a single genetically valid tree for Bantu using the comparative method, as “Bantu languages have the remarkable ability to act much more like a dialect continuum than as discrete and impermeable languages” (Schadeberg 2003:158). Conversely, Möhlig (1981) posits that the current distribution of the Bantu languages is the result of multiple layered proto-systems. In his ‘stratificational model’, Möhlig postulates as many as eleven such proto-systems for Bantu: three forest nuclei and eight savannah nuclei. In contrast to the comparative method and the stratification model, the third method that has been used to classify the Bantu languages, namely lexicostatistics, is (relatively) quick and straightforward: count the number of cognates for a limited number of basic-vocabulary items across a large number of language pairs, apply some statistical magic, and out comes your language tree. Basic vocabulary is used, as it is thought to be a language’s most stable part of the lexicon, which in turns allows for deep genetic links to be uncovered. In traditional lexicostatistics a so-called

3. Ironically, exactly one entry in Laman’s (1936) dictionary was labelled as belonging to the ‘Solongo’ dialect, namely wombo ‘brain’ on p. 1 100.
‘distance matrix’ or ‘similarity matrix’ is built in the process, where each value (i.e., ‘distance’) in the matrix reflects the percentage of shared cognates between a pair of language varieties (i.e., ‘cognition percentage’). Each variety is compared to each other variety, which enables hierarchical sub-groupings. The two extreme statistical methods used in deciding whether or not a certain variety (or cluster) belongs to an already established cluster of varieties are Nearest Neighbour (NN = the closest distance leads to grouping; this typically produces onion-type trees) vs. Furthest Neighbour (FN = the greatest distance leads to grouping; this typically results in more balanced trees). Various types of averages between these extremes can also be called in, known as Branch Average (BrAv) and Group Average (GrAv).4

Lexicostatistics found immediate appeal the moment it was first proposed in the 1950s by Swadesh (1950, 1952, 1955). According to Isidore Dyen (in Hymes 1960:35), lexicostatistics can be used: “(1) to establish the relationship between languages, (2) to classify (subgroup) related languages, and (3) to establish the times at which related languages began to diverge”. Despite a flurry of both empirical (e.g., Lees 1953) and theoretical (e.g., Sankoff 1970, 1973) work, use (3)—also known as glottochronology—has now (mostly) been discredited (e.g., Blust 2000, McMahon & McMahon 2006, Heggarty 2010). We will not be dealing with (3) in the present article, but will be looking at and working extensively with (1) and (2).

Already in 1956 the first two lexicostatistic studies for selected Bantu languages appeared in print, both by linguists from Tervuren: Coupez (1956) compared Mongo with Rwanda, and Meeussen (1956) compared Bobangi with Zulu, and both of these briefly with Ganda. During the first two decades of Bantu lexicostatistics, the studies were indeed limited in scope and essentially contrasted the basic vocabulary of just two up to a handful of Bantu varieties, for instance: Voeltz (1962) looked at Swahili and Xhosa, Obenga (1968) at Kikongo and Mbosi, Williams (1973) at the Luyia group, and Frost (1974) at Lamba, Bemba, Luba and Lunda. During those first two decades, serious attempts were also made to classify the whole of Bantu, using a so-called representative sample of Bantu languages. Examples include: Olmsted (1957), Évrard (1966), Mann (1970), Heine (1973), Henrici (1973), Coupez et al. (1975) and Heine et al. (1977).5 While Mann (1970) and Henrici (1973) basically applied proper (lexico)statistic calculations to the 28 ‘test languages’ from Guthrie (1967-71),6 in Tervuren Coupez et al. (1975) increased the sample to 57 Bantu varieties. Over the next two decades and a half, the Tervuren team would painstakingly continue to add Bantu varieties—collecting the so-called ‘Swadesh 100 list’ (pruned to 92 in their case) for each variety—reaching 214 (204 Bantu + 10 Bantoid) vocabularies in Bastin et al. (1983), and finally as many as 542 (530 Bantu + 12 Bantoid) in Bastin et al. (1999). Other attempts to classify the whole of Bantu used either smaller samples, such as Holst (1992), or looked beyond (Narrow) Bantu

4. Worked examples can be found in Schadeberg (1986:71-3) and Bastin et al. (1999:108-13).
5. Observe that parts of Heine’s (1973) and Heine et al.’s (1977) classifications are additionally supported by lexical innovations.
6. As pointed out by Hinnebusch (1999:175), Guthrie ignored lexicostatistics: “Guthrie developed his own statistical approach in his comparative study and developed several measures: a coefficient of Bantuness, indices of relationship, and modulus of dispersion (Guthrie 1967-71, Vol. 1:96-110).”
as well, such as Bennett & Sterk (1977) who studied south central Niger-Congo,7
Gerhardt (1980) who included the ‘not-so-Bantu languages’, or Piron (1995,
1997, 1998) and Bastin & Piron (1999) who considered both Bantu and Bantoid.
The remainder of the Bantu lexicostatistic studies have either dealt with particular
countries or parts of them: Uganda (Ladefoged et al. 1972), Cameroon (Dieu et
al. 1976, Nseme 1979, Hedinger & Ekandjoum 1980), Zambia (Kashoki & Mann
(Bouka & Ndamba 1991), Botswana (Batibo 1998), Cameroon and Equatorial
Guinea (Chumbow et al. 2007), etc.; or studied particular Guthrie zones or groups:
zone S (Finlayson 1976), groups A40-A50 (Schadeberg & Voorhoeve 1979), group
B30 (Van der Veen 1991), zone B (Bastin et al. 1994), zone F (Masele 2001, Masele
& Nurse 2003), groups B10-B30 (Mouguiama-Daouda & Van der Veen 2005), etc.;
or considered specific geographic areas: eastern Bantu (Hinnebusch 1976a, Nurse
1978), southern Bantu (Tunks 1993), the Great Lakes (Schoenbrun 1997), etc.; or
looked at particular language and dialect groups: Shona (Kinsman 1975, Ehret
& Kinsman 1981), Swahili (Hinnebusch 1976b, Bastin et al. 1989, Batibo 1989),
Chaga (Nurse 1977, 1979), Sabaki (Nurse & Hinnebusch 1993:271-85), Nguni
(Miti 1996), Tswana (Batibo 1999), Kikongo (Ntunda Nzeza 2007), Kiyaka (Ntota
Maselo 2014), etc. Still other publications are of a more contemplative nature, and
include: Bennett (1976), Gregersen (1976), Mann (1980), Nurse (1982), Vansina
and Mann (1999).

Whereas all of the previous Bantu lexicostatistic studies were lexical in nature,
statistical classifications based on grammatical features have also been undertaken.
This work was the brainchild of the Tervuren linguist Yvonne Bastin, who wrote
her M.A. dissertation on the topic (Bastin 1977, reprinted as Bastin 1983, and
summarised as Bastin 1980). In her study, Bastin tabulated a total of 52 phonological
and morphological features for 80 Bantu languages. A cross-comparison of the
resulting Bantu trees obtained via lexical vs. grammatical statistics is undertaken
in Bastin (1979) and Bastin et al. (1979). In the latter, 69 Bantu varieties are kept
which are common to both trees, with 92 features on the lexical tree, and 52 on
the grammatical one. Unsurprisingly, the branching on the trees agrees in very
broad terms only, not in any detail. Bantu statistical classifications that combine
both lexical and grammatical features are actually not that common. Two recent
examples are: de Luna (2010) for Bantu Botatwe (23 varieties, 100 lexical +
some grammatical features) and Petzell & Hammarström (2013) for Greater Ruvu
(7 varieties, 500 lexical + 27 grammatical features). Also, the juxtaposition of trees
that are the outcome of a lexical distance matrix on the one hand and a grammatical

7. This classification is actually part lexicostatistics, part lexical innovations.
8. In Hinnebusch et al. (1981) the lexicostatistic aspects play only a minor role: “The
approach is that of a modified comparative method, with some utilization of lexicostatistics”
(pp. 12-3).
distance matrix on the other, are effectively rare in Bantu studies. Bastin et al. (1979:380-1) did so using GrAv trees, while Petzell & Hammarström (2013:147, 149) do so using Neighbour-Joining trees (Saitou & Nei 1987). For the latter method, see Section 1.5.2.

1.3.2. The Tervuren classifications (Bastin et al. 1999)

The Tervuren classifications of the Bantu language family found in Bastin et al. (1999) deserve a separate section, not only because they are the result of the longest-running lexicostatistic project for Bantu (Achille Meeussen took the initiative in 1953, cf. Bastin et al. (1999:iii)) or because this study included far more varieties (542) than any other lexicostatistic project, but because at least two novelties were introduced (‘composite heterograms’ and ‘Variable Neighbourliness’) and, most importantly, because the Tervuren data continues to be (re)used to this very day. In order to properly appreciate the Tervuren work and in order to be able to interpret the new uses of the data, a good understanding of their initial work and results is thus necessary. For obvious reasons, Kikongo will be used in the exemplifications below.

Starting from the Swadesh 100 list,9 eight basic-vocabulary items were dropped and a further four renamed. The dropped items are: I, we, thou; not; this, that; green, yellow;10 while the following were renamed: earth > ground; foot > leg; hand > arm; claw > nail. In Bantu, I, we, thou and not rather belong to grammar than lexis, this and that do not lend themselves to linguistic comparisons, green and yellow are not common concepts, while the renamed items are culturally more appropriate for Bantu. A questionnaire was drawn up to facilitate the collection of equivalents for the 92 concepts, for which: “Generic terms were to be preferred to specific, the grammatical category of the gloss was not binding (for example, an adjective might be rendered by a verb or noun if that was more natural in the language being collected), and multiple responses were allowed where several terms were in equal currency” (Bastin et al. 1999:5). For some varieties, material was also excerpted from published sources. The actual cognition judgements were done in Tervuren, the statistical processing by Michael Mann in London, who observed: “For multiple data, I scored a match if any root-code for language x agreed with any root-code for language y,” and: “For missing data, I counted the glosses for which data was jointly available for x and y, and expressed the matches as a percentage of this figure” (Bastin et al. 1999:9). The former is an approximation, as ongoing vocabulary replacements may thereby not be taken into consideration, while the latter may result in skewing for small sets.

For visualisation purposes, a map was produced with sample points (rather than language areas) for each of the 542 varieties; Figure 12 shows the Lower Congo region.

---

9. That list was first published in Swadesh (1955), where one needs to select the 92 ‘starred forms’ from Table 2 (pp. 133–7), to which the ‘eight additions’ mentioned on p. 124 must be added.
10. Although often repeated in the literature, including by the team members themselves (e.g., Bastin et al. 1999:v), it is thus not the case that the Tervuren team took out items like ice and snow—those had already been taken out by Swadesh.
As may be seen from Figure 12, no attempt was made to differentiate between languages and dialects, and for some varieties multiple regiolects were included (e.g., H16k/1, H16k/2, etc.). Each Bantu variety then becomes just that: a sample point on the map. For Kikongo, obvious varieties are missing, such as Kisolongo and Ciwoyo, both spoken at the mouth of the Congo River, while the placement of others seems suspect: H16i/3 Sondi is claimed to be spoken in Mavunda Ndemvu, which is in the Bandundu province of the DRC (circled in Figure 12). Nonetheless, the similarity with the geographically distant H16jH Hangala and H16jK Kamba, both in Congo-Brazzaville, is as high as 89%, and with the even further H16i/2 Sundi from Mboko-Songho, also in Congo-Brazzaville, it is 88% (Bastin et al. 1999: 102). This Sondi is thus indeed Kikongo, but was likely mapped wrongly. Contrariwise, B83/1 Mfinu and B83/2 Mfinu (Kinsi dialect), both in the DRC and fully surrounded by H16 varieties (cf. Figure 12), are a total mystery. These varieties were seemingly transcribed by Jan Daeleman in 1980 (Bastin et al. 1999:13), but are clearly not Kikongo: the similarities with the surrounding H16 varieties vary from 38% to maximum 52% (Bastin et al. 1999: 101).

The first novelty introduced by Mann is the set of so-called composite heterogloss maps, drawn on stylised maps with polygonal boundaries around each language, whereby the series of 92 isoglosses for each variety are combined and compared with those of the surrounding varieties, “some 1 611 measurements in all” (Mann 1999:167). These comparisons may be carried out at different similarity levels, and effectively result in maps at any chosen similarity percentage, best

Figure 12. Sampled language varieties in the Lower Congo region in Bastin et al. (1999:6; the highlighting of the H10 and B40 varieties was added by us).
viewed “on a computer screen, with a slider permitting the similarity threshold to be continuously adjusted, and a facility to zoom in on any area of interest” (Bastin et al. 1999: 76, note 3). As an illustration, Figure 13 shows the 65% composite heterogloss map for the Lower Congo.

Figure 13. Composite heterogloss map for the Lower Congo, with links at or above 65% similarity dotted in grey and bounds below 65% similarity in black (Bastin et al. 1999:84; the colouring is ours).

Continuities and discontinuities may easily be read from such maps. As expected, the already mentioned B83/1 and B83/2 are fully isolated at this level of similarity, but surprisingly, so are H12b/2 (a Kiyombi variety from Congo-Brazzaville, spoken in Kakamoeka) and H13b (a Sundi variety from Congo-Brazzaville, spoken in Kifouma). This particular map was chosen for the fact that at this threshold, 65%, the Kikongo group ‘breaks through’ into the B40 group. Lowering the threshold further, to 60%, results in even more B40 languages joining the Kikongo group in the north, but at the same time ‘leaks’ appear in the south, with also H30 and B80 languages joining the Kikongo group, and through another leak in the north also B50-60-70 languages. The latter already started at the 65% level, as may be seen from Figure 13, where the leak language is H11/2, i.e., Bembe as spoken in Mabombo, Congo-Brazzaville. For some reason, this particular use of lexicostatistics to obtain geographical classifications “reflecting areal influences” (Mann 1999:165) has not received any further attention in the literature. Especially Mann’s ‘slider’ seems like a very worthwhile application to be developed in this day and age of digital ubiquity.

Conversely, hierarchical classifications “inviting historical interpretation” (Mann 1999:165) have continued to receive attention. Being dissatisfied with the
then standard methods in lexicostatistics to cluster languages—NN, FN, BrAv and GrAv—Mann decided to produce a graded series of similarity trees by requiring that a sixth, a third, a half, two-thirds or five-sixths of the similarity levels in the material to be grouped be above the current similarity level. He coined his method *Variable Neighbourliness* (VN), abbreviated to respectively VN17, VN33, VN50, VN67 and VN83. The classic NN (Nearest Neighbour) and FN (Furthest Neighbour) are then conformably abbreviated to VN0 and VN100. Rather than end up with a single tree for the whole of Bantu, as many as seven Bantu-wide trees were thus produced, to which the GrAv and BrAv trees were added for good measure. These nine Bantu-wide trees are therefore all derived from the same distance matrix of 146 611 pair-wise comparisons (which is the figure one obtains when comparing each variety with each other variety in the set of 542); the difference lies in the statistics. The stated beauty of the novel VN\(x\) approach is that trees with low values of \(x\) will tend to reflect lateral influences (i.e., diffusion), while trees with high values of \(x\) will rather point to historical (i.e., genetic) links (as reported in Bastin & Piron 1999:150, based on p.c. by M. Mann). Mann further defined two important concepts, *connectivity* and *exclusivity*: “Connectivity measures the completeness of intra-group internal relationship before the group is absorbed into some wider group [...] exclusivity measures the incidence of extra-group similarities when the group is formed or at other defined levels, e.g. at the level of the most distant intra-group link” (Bastin *et al.* 1999:105). In the various published trees, these two measures are presented graphically at each node with two half-pies filled in black from right to left: the top one for exclusivity, the bottom one for connectivity. As an illustration, the H10 section in the VN33 tree, which is “biased towards local relationships while avoiding extremes” (Bastin *et al.* 1999:194), is as shown in Figure 14.

Despite the wealth of data summarised in Figure 14, and despite the fact that as many as 22 Kikongo varieties were included in this Bantu-wide study, the overall internal classification of the Kikongo varieties remains disappointing.

The data seen in Figure 14 basically separates the Kikongo varieties spoken in Congo-Brazzaville from the others.\(^{11}\) What is apparent from the detailed statistics, however, is the close relationship of H10 with B40. Although Figure 14 is how the Tervuren team decided to present their ‘H1 Kongo group’, a better view is actually to consider the full clade to which these branches belong, as done in summarised form in Figure 15, where the last two nodes are the summary of Figure 14.

---

11. With H16i/3 Sondi mis-mapped, as argued above.
Figure 14. Kongo group in Bastin et al.’s (1999:214-5) VN33 tree, followed by the details for all other trees.  

Figure 15 is a much more satisfactory representation: This one clade, which we would like to call the ‘Kikongo clade’, is coordinate with a clade which basically consists of B50-B60-B70-B80 languages (not shown here). The Kikongo clade in the VN33 tree, then, suggests that the H10 languages belong to a group that also includes a sub-group with B40 languages, as well as a sub-group with B30 languages and B85/7 Yanz (Nsambaam dialect). In all fairness, it must be pointed out that two more H10 varieties, H1/1 Bembe Mouyondzi and H1/2 Bembe Mabombo, are classified as belonging to the coordinate clade (of B50-B60-B70-B80 languages).
languages, which also includes H24 Songu; not shown here), while H13b Sundi Kifouma is assigned to a much earlier clade in the VN33 tree (of languages from zones K, L and R, together with H43 Sondi in the Bandundu; not shown here either).

**Figure 15.** Extract from the VN33 ‘summary tree’ from Bastin et al. (1999:139), with a focus on the ‘Kikongo clade’ (our term). Black triangles indicate the similarity levels of the summarised nodes; the colour codes have been added by us.

**Figure 16.** Extract from the VN33 map from Bastin et al. (1999:140), with a focus on the ‘Kikongo clade’ (our term). The colour codes have been added by us, and reflect the sub-groups within the Kikongo clade, as seen from the summarised tree in Figure 15.

Implicit in the Tervuren data, therefore, is the following *internal classification* of the Kikongo clade: There is an initial split between Civili, Kiyombi and B40 languages on the one hand, and H10-H30-H40 languages on the other. The first group then splits into three sub-groups: (a) H12a/2 Civili in Gabon, (b) B40 (Lumbu
and Bwisi), H12a/1 Civili in Congo-Brazzaville and H12b/2 Kiyombi, and (c) B40 (Shira-Punu). The second group splits into four sub-groups: (a) H30-H40 (Kiyaka, Kisuku, Hangana) and B85/7 Nsambaan, (b) H13a Kikunyi, (c) H16 languages south of the Congo River and H12b Kiyombe, and (d) H16 languages north of the Congo River. These sub-groups may be visualised as in Figure 16.

1.4. Lexicostatistics for the Kikongo varieties

Although extracted from a Bantu-wide classification, the Kikongo clade in the VN33 tree from Tervuren seen in Figure 15 and mapped in Figure 16 is the only one for Kikongo in the existing literature which does not stop at present-day country boundaries. Country studies of languages in sub-Saharan Africa, while highly revealing in themselves, will always result in truncated views. In their lexicostatistic study of the languages of Congo-Brazzaville, for example, Bouka & Ndamba (1991) end up with a sub-group G2A (using the BrAv method) which brings together B40, B50, B73b,c (Teke varieties) and H10. Without the bigger picture, i.e., without varieties from across their country’s boundaries, the classification of any of their language groups cannot be complete. The fact that B40 ends up in a clade together with H10 is thus only indicative in this study (and in any case, B50 ends up in the same group, as does Teke).

Conversely, restricting the study of the Kikongo varieties to those spoken in the DRC only has the potential to result in an internal classification close to reality, given the central position of the DRC in the Kikongo clade, whereby all the DRC border varieties at least hint at the varieties across the border. A closer look at the lexicostatistic study of Ntunda Nzeza (2007), who classified the Kikongo varieties spoken in the DRC, is thus in order. On the face of it, Ntunda Nzeza did everything right: working with a massive 208 consultants (most through correspondence, however), he collected the Tervuren 92 equivalents in as many as 198 localities spanning the whole of the Lower Congo province of the DRC. He then claims (Ntunda Nzeza 2007:64) to have used 33 of the 92 basic-vocabulary items to draw isoglosses. This then supposedly led to the delineation of 25 dialects or, including sub-dialectal forms, 36 Kikongo varieties. This may or may not actually have been done; the evidence is not provided. Rather, all of this is preceded by an informative discussion of 26 Kikongo glossonyms, all bar one (Kimanyanga) which are then used to name his 25 Kikongo dialects. Also note this important statement at the start of the section on the glossonyms: “Le regroupement d’individus autour d’un lignage, d’un clan ou encore sur un lieu d’origine bien déterminé a servi, pour nous, de critère important pour la classification des variantes dialectales koongo. La plupart des dialectes sont désignés par les termes des clans ou des lignages auxquels appartiennent respectivement leurs locuteurs” (Ntunda Nzeza 2007:36). In Figure 17 we have depicted the main localities of each of Ntunda Nzeza’s 36 varieties on a map showing the districts, territories and collectivities of the DRC’s Lower Congo province.
Figure 17. The 36 Kikongo varieties in the DRC’s Lower Congo province according to Ntunda Nzeza (based on Ntunda Nzeza (2007:62); the colouring of the codes is ours).

As it turns out, Ntunda Nzeza’s 36 Kikongo varieties from the DRC and their associated isoglosses typically correspond to the borders of either one or two, sometimes more, collectivities. Ntunda Nzeza does not show us any actual language data, so we have to take his distance matrix for granted. Furthermore, using some idiosyncratic statistical approach he arrives at a perfectly symmetrical tree which consists of twofold splits at each level. For readability, we have redrawn that tree, supplementing the codes with the language names derived from the data in his study (combining glossonyms and, where needed, localities). This tree is shown in Figure 18. According to Ntunda Nzeza, the Kikongo dialects ‘fragment’ into two main blocks (A and B), each of which splits into two (1 and 2), then again into two (α and β), and then once more into two—with the end result being 16 distinct sub-groups of two, occasionally three, varieties (1 to 16). He then argues (Ntunda Nzeza 2007:106) that block A consists of geographically peripheral Kikongo varieties (in the DRC), while those from block B consist of geographically central varieties. He also posits that the appearance of those two blocks is a recent phenomenon, stemming from the end of the 19th century (cf. quote in the introduction), and lists a number of recent psychological, cultural and matrimonial traits of the centre vs. the periphery in further support.
Figure 18. Tree of the 36 Kikongo varieties in the DRC’s Lower Congo Province according to Ntunda Nzeza (based on Ntunda Nzeza 2007:104); the colouring of the varieties is ours.
In his general conclusion, however, Ntunda Nzeza also suggests quite the opposite, hypothesising that each variety contains traces of the original populations, subjugated by the Kongo: “S’il faut déterminer le statut linguistique de ces 36 parlers nous dirons [...] que ces 36 dialectes koongo en question seraient des vestiges linguistiques que les peuples Teke, Wumbu, Mbundu... auraient gardés après avoir adopté la langue des conquérants” (Ntunda Nzeza 2007:110). This refers back to a process that started at the end of the 13th century (Ntunda Nzeza 2007:14). No evidence is presented in support of this hypothesis; unlike for the end-of-the-19th-century ‘fragmentation’ hypothesis.

1.5. Phylogenetic classifications of the Bantu languages

Lexicostatistic trees are based on similarity levels, while one would actually want to build trees based on shared innovations only (François 2014). In that, lexicostatistics results in phenetic (observational) trees—i.e., according to appearances (phenomena)—and not cladistic (evolutionary) trees. In the Bantu context, this important difference—and the failure to pursue the second approach—was for example pointed out by Flight (1988). Michael Mann was aware of this problem (Bastin et al. 1999:2-3), but after nearly half a century of work, they must have concluded that it was better to publish their results ‘as is’, rather than to re-launch themselves. Additionally, as Atkinson & Gray point out in their article on curious parallels and connections between the development of methods in biology and historical linguistics:

Despite the methodological similarities between the comparative method in linguistics and biological cladistics, up until very recently, historical linguists did not typically use computer algorithms to search for the best language tree(s). This is surprising given that the task of finding optimal trees for even a moderate number of languages is one of considerable computational complexity [...] (Atkinson & Gray 2005:518)

The necessary computer power became available at the turn of the millennium (Pagel 2000), and despite some initial ‘interdisciplinary indiscipline’ (Heggarty 2006), the first few cladistic studies for Bantu using advanced computational algorithms soon followed. The types of classifications built in this new era are known as phylogenetic, with phylogeny defined as: “A graph depicting the descent of a set of individual units. The units are typically word lists or DNA molecules. The term encompasses both phylogenetic trees and phylogenetic networks” (Forster & Renfrew 2006:197). It is important to stress that this particular use of ‘phylogenetic’ is peculiar to linguistics. As Dunn points out in his recent overview of linguistic phylogenies: “In principle, all historical linguistics is phylogenetic [...] The reason that in linguistics this term is often used in contrast to other forms of historical linguistic investigation is that phylogenetic approaches maintain their methodological link to the investigation of evolutionary processes in other, mostly biological, domains” (Dunn 2014:190).
1.5.1. The Tervuren data reused

While the Bantu-wide classifications of Bastin et al. (1999) concluded nearly half a century of lexicostatistic research at Tervuren (Yvonne Bastin retired in October 2003), the underlying lexical data, as well as the grammatical data from Bastin (1977, 1983), has proven to be a valuable starting point (and goldmine) for a series of new types of Bantu classificatory studies.13

1.5.1.1. Lexical reuse

Mindful of Guthrie’s (1959:73) observation, quoted in the introduction, that classifications are typically not an end in itself but are rather built with other goals in mind, it needs to be pointed out that this is also the case for all but one of the Bantu phylogenetic studies published to date. This is even apparent from the title of the first, trailblazing article from this new era of Bantu classifications, namely “Bantu language trees reflect the spread of farming across sub-Saharan Africa” (Holden 2002). Here, a novel language tree for Bantu is generated on the one hand, while dated archaeological information on farming in sub-Saharan Africa is plotted on a map on the other. Cross-comparing those, Holden concludes that modern Bantu sub-groups, defined as clades in her trees, mirror the first farming traditions.14

In other words, a language tree is used to model population history, which is the most straightforward use indeed. The subtitle of her article tells us something about the methodology used, namely “A maximum parsimony analysis”, which may be defined and characterised as follows:

Maximum Parsimony = Procedure for reconstructing the shortest possible (i.e. most parsimonious) tree from the data at hand. The term ‘shortest tree’ refers to a tree with the minimum number of changes along its branches. The method is applied to character data […] (Forster & Renfrew 2006:196)

Maximum-parsimony language trees have several advantages compared with either traditional linguistic trees or lexicostatistical trees. Parsimony methods use a computer-implemented algorithm to search for the best tree according to an explicit optimality criterion […] The resulting tree(s) is bifurcating,15 with branch lengths proportional to the number of linguistic innovations. Unlike lexicostatistical trees, which are based on a measure of overall similarity among languages, parsimony methods only use derived characters or shared innovations to define descent groups. (Holden & Mace 2003:2431)

13. While the focus of Section 1.5 is on phylogenetic classifications of the Bantu languages, it must be pointed out for completeness that the Tervuren data sets have been reused in still other classificatory approaches as well. For example, Wotzka & Bostoen (2009) have presented a lexical classification of zone C by correspondence analysis, drawing all their data from Bastin et al. (1999).
14. The actual correlation is not proven, however; Holden simply assumes they truly mirror one another, as the successive clades in her tree ‘look like’ the successive archaeological traditions on her map. Ironically, early Bantu speakers were not farmers (cf. Bostoen forthcoming).
15. One adjudicator pointed out that it is not inherent in the technique of maximum parsimony that only bifurcating trees can be produced.
The latter, character data (or discrete data), as opposed to distance data, is the important difference. In traditional lexicostatistics, all that remains per language pair is a single blunt similarity measure, $x\%$. In contrast, when using character data, the actual realization of each character is kept track of. For instance, the state for the basic-vocabulary character ‘to bite’ is -táčika in Kisikongo, -lunza in DRC Kisolongo, -tatika in Kizombo, and -tebila in Kiyombi. This may then be converted into a single column of multistate data, here: 1, 2, 1, 3.

In her study, Holden (2002) constructed a tree for 73 Bantu plus two Bantoid varieties—the latter two outgroups are needed to ‘root’ the tree. These 75 varieties were selected from Bastin et al. (1999), not on linguistic nor distributional grounds, but based on the availability of cross-comparative (here cultural) data. From a linguistic point of view, this selection is thus rather random. The cognition judgements for each character state were also taken from Bastin et al. (1999). Vocabulary items with two forms in a single language variety (i.e., synonyms) were treated as polymorphic character states. The resulting tree, perhaps disappointingly, largely corresponds with the earlier lexicostatistic ones from Bastin et al. (1999). The Bantu-wide view is illustrated in Figure 19.

![Figure 19](image)

**Figure 19.** Mapped sub-groups of Holden’s (2002) maximum-parsimony Bantu tree (coloured ellipses and numbers for the sub-groups, with those numbers reflecting the progression through the tree, were added by Grollemund (2012:160)).

Specifically for Kikongo, a bootstrap analysis of RC (i.e., rescaled consistency index) reweighted Bantu lexical characters, retaining clades in 50% or more bootstrap pseudo-replicates only, showed that H10 and H30 form a clade together with B73 Teke. There were no B40 varieties in the sample, however, which makes further
analysis on this micro-level unhelpful for our purposes. The overall conclusion is potentially important: “The correlation between archaeology and language groups suggests that the major subgroups of modern Bantu [i.e., those shown in Figure 19] stem from the Neolithic and EIA, with little subsequent movement by speech communities. It also suggests that the modern languages within each subgroup evolved in situ in these areas” (Holden 2002:799).

As an illustration of the power of a language tree built according to the (then) latest models, Holden and her colleagues went on to put their tree to good use in a number of additional studies, showing that phylogenetic methods for inferring ancestral cultural states can lead to both plausible and falsifiable results (Holden & Mace 2003, 2005). They also introduced maximum-likelihood (ML) tests for correlated evolution among discrete binary characters on a Bantu tree, in their case testing for coevolution between cattle and matriliney, and cattle and patriliney.¹⁶

A decade ago Holden and her colleagues then moved to the next step, by introducing Bayesian tree sampling for Bantu. As such, Holden et al. (2005) may be seen as a methodological article, in the sense that Bantu trees are built and compared for the sake of building trees—in an attempt to arrive at the ‘true tree’. In simple terms:

Bayesian methods for inferring phylogeny do not search for the best tree(s) according to an optimality criterion such as maximum parsimony or maximum likelihood, but instead sample a large number of trees (from the universe of possible trees) in proportion to their likelihood. Trees with a higher likelihood will be sampled proportionately more often, but trees with an intermediate likelihood, which occur more frequently in the universe of possible trees, will also be represented in the sample. The aim is to represent phylogenetic uncertainty within the tree sample. The proportion of trees in the sample with a given node is equivalent to the Bayesian posterior probability of that node [...] (Holden et al. 2005:60-1)

To compute the posterior probability distribution of trees built with Bayesian statistics, one typically uses Markov chain Monte Carlo (MCMC) algorithms. Also, instead of multistate data, Bayesian methods use binary data: one column of 0’s (i.e., absence) and 1’s (i.e., presence) per cognate, which “has the useful property that it can handle more than one cognate set per meaning” (Dunn 2014:191). Or thus for the above example, three columns: 1010, 0100, 0001—each additional cognate is simply an additional column. It is clear that this method of encoding the cognates produces dependencies between the columns (Evans et al. 2006:124). While this at first seems to be problematic, Pagel & Meade (2006:181) have shown that this “non-independence should have only a small effect on the best topology that is retrieved from an MCMC sample”.

The cognate data from Bastin et al. (1999) was again reused, but the number of varieties raised from 75 to 95 for the Bayesian trees. When comparing the maximum

¹⁶. For which results showed that adopting cattle led to the loss of matrilineal descent and its replacement by either patrilineal or mixed descent, but this without conclusive evidence that early Bantu-speaking populations were matrilineal, while the earliest Bantu-speaking populations did lack cattle, and that cattle-keeping was adopted at least twice in east and south-east Africa.
parsimony trees (73+2 varieties) with the Bayesian tree sample (93+2 varieties), Holden et al. (2005) conclude that the trees concur on several levels, with some (but not all) uncertainties resolved. In contrast to the maximum parsimony trees, for instance, “Bayesian trees strongly support the view that South West Bantu is co-ordinate with East Bantu, rather than clustering with languages spoken in the rainforest” (Holden et al. 2005:62). As to Kikongo, there were no B40 varieties in the sample of 95 either.

In Holden & Gray (2006) yet another level is introduced in an attempt to eliminate all remaining uncertainties in the Bantu-wide classifications. That added level is a Neighbor-Net, but before doing so Holden & Gray also offer a new ‘majority rule tree summarising a Bayesian sample of trees’ in which an initial split between West Bantu and East Bantu is now apparent (cf. also the discussion in the introduction), and this despite the fact that the same set of 95 varieties as in the previous study (Holden et al. 2005) was used. In simplified terms, Neighbor-Net may be characterised as follows:

Neighbor-Net (Bryant & Moulton 2004) is an agglomerative method for constructing networks that selects taxa on the basis of similarity and groups them together. [...] In agglomerative tree-building methods, two taxa (or nodes) are chosen on the basis of similarity, then they are agglomerated (merged), the data matrix is reduced and we proceed to the next iteration. However, to construct a network, the Neighbor-Net method does not immediately agglomerate the selected taxa (or nodes). Instead, it waits until one of the chosen taxa (or nodes) has been grouped with a different node. Then the three nodes are reduced to two, and the process is repeated. The Neighbor-Net method represents similarities within the data set as a splits graph, constructed from a distance matrix. Splits are bipartitions of the data. (Holden & Gray 2006:25)

Observe thus that networks in Neighbor-Net are calculated from distance data. Specifically for West Bantu, the Neighbor-Net network shown in Figure 20 is offered. Examining their Neighbor-Net, Holden & Gray (2006:29) conclude that there was a more or less simultaneous divergence of six sub-groups of West Bantu languages, one of them the H languages plus B73 Teke and B80 Madzing. (Recall that there were no B40 languages in this sample, however.) Their overall conclusion for this region is that: “West Bantu languages underwent a rapid early radiation—near simultaneous divergence of several major West Bantu branches—resulting in a star-shaped phylogeney with little internal structure. With the exception of the isolated language Bubi A31 (and possibly Mpongwe B11) we found no evidence that the greatest divergence within Bantu involved languages of zones A and B. There is little evidence for borrowing among West Bantu languages, possibly indicating that these speech communities were more isolated from one another than were early East Bantu-speaking populations” (Holden & Gray 2006:29).
Figure 20. Network for a sample of 29 West Bantu languages (plus the Bantoid Tiv and Ejagham as outgroups), calculated using the Neighbour-Net method (Holden & Gray 2006:27). The colouring, which we added, shows Holden & Gray’s six sub-groups for West Bantu.

In contrast, and this even though the same dataset of 93 + 2 varieties was again put to work, the Bantu-wide tree of Pagel & Meade (2006:176), which is derived from a Markov chain, does uncover a homeland of A and B languages. These languages are coordinate with everything else, with the latter consisting of two coordinate clades: West vs. East. The authors do not discuss the implications of this new arrangement at all, though, given that their main concern is the introduction of yet another novelty for Bantu phylogenetic trees, i.e., the estimation of the rates of lexical replacement on those trees.\(^{17}\)

In 2010, the 93 + 2 dataset was used once again, by Walker & Hamilton (2010), in their study of the rise and fall of social complexity, but pruned to 87 + 2 varieties as they needed corresponding data for ‘jurisdictional hierarchy beyond local community’. The Bantu-wide Bayesian MCMC phylogeny they present (Walker &

\(^{17}\) In follow-up work the same authors, together with colleagues, showed that: “31% of vocabulary differences among Bantu language speakers arose at or around the time of language-splitting events” (Atkinson et al. 2008:588), meaning that a third of the Bantu basic-vocabulary items arose in rapid evolutionary bursts from the predecessor tongue. Given that such rates are not the focus of the present article, nothing more will be said about them. We will, however, do so in a forthcoming article. On topic: alas, no actual trees are shown in Atkinson et al. (2008), not even in the supplementary material, but one may assume the tree to be similar to the one in Pagel & Meade (2006:176).
Hamilton 2010:1402) now contains a ‘North and Forest West Bantu’ procession of languages, with Kikongo & Co. surrounded by languages from zones A, B and C. Another phylogenetic Bantu tree which is presented too briefly to be of much value for our purpose is the one found in Dunn et al. (2011:80)—a so-called maximum clade credibility tree, for 21 varieties + 1 outgroup. Their Bantu tree has no languages from zone H, nor from zones A, B, F and M for that matter.

The most recent published study in which the Tervuren data is reused is the one by Opie et al. (2014), who apply Bayesian phylogenetic comparative methods to infer the ancestral state and evolutionary trajectory of Bantu kinship patterns. In the supplementary material to their study they present maximum clade credibility trees consisting of 122 language varieties (119 Bantu + 3 Bantoid). With regard to Kikongo, their data show that Yombe, Sikongo and Sundi are coordinate with Suku and two Yaka varieties.

1.5.1.2. Lexicogrammatical reuse

Rexová et al. (2006) is the only Bantu-wide phylogenetic study in which both lexical and grammatical data from Tervuren are reused: a total of 87 varieties were selected and 144 characters taken into account (the 92 basic-vocabulary items, plus 52 grammatical features). After generating both maximum parsimony and Bayesian trees, the authors conclude that: “The main phylogenetic signal of our data favors the colonization of Angola, SW Congo-Kinshasa and surrounding territories from the more eastern source areas” (Rexová et al. 2006:191). Unfortunately, also this selection did not contain any B40 languages.

1.5.1.3. Phylogeographical reuse

One had to wait until 2013 to see the first reuse of the full lexical dataset from Bastin et al. (1999). Currie et al. (2013) undertake to build a Bayesian MCMC tree, the structure of which they then map in geographical space in order to contribute to the debate on the pathway of the Bantu expansion. They encountered two problems with the published Tervuren data, however, which forced them to lower the number of vocabulary items used and to lower the number of varieties that could be mapped. Firstly, they noticed that the published cognation judgements for *nose* and *one* (Bastin et al. 1999:54) were repeated on the page where one would have expected those for *name* and *neck* (Bastin et al. 1999:52). In order to avoid duplication, this means that anyone who is reusing the published cognation judgements for *nose* and *one* must be aware that the published cognation judgements for *name* and *neck* refer to a different set of items.

18. As to their conclusion, the dynamics of the Bantu expansion were “wave-like exhibiting gradual increases and decreases through time, with an overall net increase in social complexity over the expansion phase” (Walker & Hamilton 2010:1403).
19. Dunn et al. (2011) studied word order across language families and conclude with: “cultural evolution is the primary factor that determines linguistic structure, with the current state of a linguistic system shaping and constraining future states” (p. 79).
20. According to the authors of this study, the reconstruction of the ancestral state of Bantu kinship reveals that inheritance and residence rules coevolved as farming spread throughout sub-Saharan Africa. Observe that the authors did not really build a new tree, as they extracted language varieties from the tree by Currie et al. (2013)—for which see Section 1.5.1.3.
judgements of Bastin et al. (1999), actually only has 90 vocabulary items available. This immediately begs the question: How did all the other researchers who reused the Tervuren lexical data handle this problem? If they did nothing, which is likely, they all introduced an error right from the start! Secondly, when Currie et al. tried to map all 542 varieties, using the longitude and latitude data provided in Bastin et al. (1999:9-23), as many as 35 (6.5%) fell outside the range indicated by the Global Mapping International language maps, with led Currie et al. to suggest that this was due to ‘opportunistic sampling’. They were therefore forced to reduce their ‘full set’ after all, to 507 varieties (495 Bantu + 12 Bantoid), and to prune their phylogenetic trees. Given this state of affairs, the idiosyncratic mappings of H16i/3, B83/1 and B83/2 discussed in Section 1.3.2. (cf. Figure 12 and Figure 13) may not surprise anymore. Currie et al.’s supplementary material indicates that H16i/3 was one of the deleted varieties, but B83/1 and B83/2 were kept. This, in turn, leads us to question the location of at least some of the 507 varieties retained—which is not a minor problem for a study in phylogeography!

Those problems aside, their results show no evidence for an early deep split between East and West Bantu. They rather find evidence for a single ‘pathway through the rainforest’, after which one branch went further south and west, another towards the Great Lakes and then further east and south-east. As far as Kikongo is concerned, a solid clade appears bringing together H10-H30-H40 and B40 languages. Only H13b Sundi Kifouma in Congo-Brazzaville and H43 Sondi in the Bandundu join the coordinate branch, where they are surrounded by K and L languages. This, of course, is reminiscent of what was seen in the Tervuren VN33 tree.

1.5.2. New data samples

After more than a decade of uses and reuses of the Tervuren data, phylogenetic classifications of Bantu languages that employ (partly) new datasets have begun to appear. In her detailed analysis of the North-West Bantu languages, Grollemund (2012) looked at 100 items of basic vocabulary across 166 Bantu varieties (from zones A, B, C, D, and groups H10-H20-H30, L10), 40 Bantoid varieties, plus Mbenbe (Jukunoid) as outgroup. The 100 vocabulary items were an extract from the 159 in the ALGAB (i.e., Atlas Linguistique du Gabon) database. The data for the 207 varieties themselves were extracted from a wide variety of published as well as unpublished sources. The cognation judgements, encoding and processing were all undertaken by Grollemund (2012:283-324). In her dataset, the B40 languages are obviously well represented; but this time it’s the H10 group which is underrepresented, with just five varieties: H11 Kibembe, H12 Civili in Gabon, H16 East Kikongo, H16c Kiyombe, and H16f Cilaadi.
In analysing her data, Grollemund (2012:325-97) uses two distance-based phylogenetic methods: Neighbour-Joining to draw trees (Saitou & Nei 1987) and Neighbour-Net to draw networks (Bryant & Moulton 2004). Neighbour-Joining may be described as follows:

A Neighbour-Joining tree is the simplest principled way to derive a tree with branch-lengths from a distance/similarity matrix. First, note that any given tree defines a distance matrix as the distance along the branches between any two leaves. Neighbour-Joining is designed to find the tree with branch-lengths whose distance matrix is as close as possible to the input matrix. The sought after tree is built iteratively by “joining neighbours” (for details see Felsenstein 2004:166-70), but it is not necessarily the closest pair of leaves (i.e. the pair with the highest cell value in the matrix) that are joined in each step. (Petzell & Hammarström 2013:146-7)

Her simplified Neighbour-Joining tree (Grollemund 2012:401), where branch lengths are not proportional to the Neighbour-Joining branch lengths, is shown in Figure 21.

![Figure 21. Simplified Neighbour-Joining tree for 166 Bantu varieties, from a study focusing on the North-West Bantu languages and the relation with their neighbours (Grollemund 2012:401).](image)

In this study, the group to which the Kikongo varieties belong is coordinate with L10-H40 languages, and both of these clades are in turn coordinate with B80-C60-C70-C80 languages. Figure 22 shows the detail for the group to which the Kikongo varieties belong, from which one sees that H12 Civili in Congo and H16c Kiyombe end up in the Shira-Punu group, while H11 Kibembe, H16 East Kikongo and H16f Cilaadi together with H31 Yaka and H42 Hungan form a sub-group on their own. As pointed out, there are unfortunately not enough Kikongo varieties to study their internal sub-grouping any further.\footnote{This is also the case for the simplified Bayesian consensus tree presented for the North-West Bantu languages in Bostoen \textit{et al.} (2015). This tree builds on the earlier Grollemund (2012); with comparable results.}
2. Our phylogenetic classification: Materials and methods

A phylogenetic classification (in the computational sense; cf. Section 1.5) of the Kikongo varieties is non-existent at present, hence our study.

2.1. Materials

As far as the raw data goes, a potentially ideal starting pack could have been the set of wordlists collected (but not published) by Ntunda Nzeza (2007) for 36 Kikongo varieties spoken in the DRC’s Lower Congo province (cf. Section 1.4). However, despite a number of efforts to get hold of those lists, the data proved impossible to obtain. As part of the linguistic fieldwork of the KongoKing research group in the DRC’s Lower Congo province, from August to September 2012, it was therefore decided to include the collection of basic vocabulary for each of the Kikongo varieties studied. The fieldwork was multipronged, with in addition to lexical data (basic vocabulary, cultural vocabulary, hundreds of reflexes of PB forms, etc.), also morphological data (especially on tense/aspect systems) and syntactic data (especially on information structure) being collected. For each variety, naturally occurring speech (i.e., oral corpus data) was also recorded and transcribed. The actual choice of the varieties studied was thus in function of these various goals, and not merely with the phylogenetic classification in mind.22 Following a review of the existing literature, it was decided to focus on three regions in the DRC’s Lower Congo: the east, the mouth of the Congo River and the north-west.

22. Over and above the study on which we report in the present article, linguistic studies that have resulted from the 2012 fieldwork so far include De Neef (2013), Dom (2013a, 2013b), Drieghe (2013), Bostoen et al. (2013), Bostoen & de Schryver (2015) and De Kind et al. (2015).
Specifically for the phylogenetic study, it was decided to collect the 92 ‘Tervuren concepts’, which implies that our data may be supplemented with the neighbouring varieties found in Bastin et al. (1999), or with those from any other Bantu study where the same list of basic vocabulary was used as a starting point. This interoperability effectively meant that there was no need for us to collect everything from scratch, so that we could instead focus on filling gaps in the documentation. Compared to the Kikongo varieties from Congo-Brazzaville included in Bastin et al. (1999), for instance, those from the DRC are particularly underrepresented. For the east of the DRC’s Lower Congo we decided to investigate the poorly documented Kimbata, Kimpangu, Kimbeko and Kinkanu. With no mouth-of-the-Congo River varieties at all included in Bastin et al. (1999), we further focussed on adding DRC Kisolongo and Angolan Kisolongo (as spoken in Mbanza Soyo) as well as Ciwoyo, and in the process also ‘discovered’ Cizali and Cimbala. For the north-west, we were in search of Laman’s “Nd = Ndingi or Ndinzi” (cf. Section 1.2.1, Figure 5) and ended up ‘discovering’ Cilinji and Kizobe. To collect the basic vocabulary, we worked with one main consultant per variety, often backed up by a crowd around him or her, presenting each of the 92 concepts to them in various syntactical contexts. Figure 1 indicates the locations from where the respondents hailed for the Kikongo varieties just mentioned. The online Supplementary material also includes the names of the respondents and further details.

In 2013 the materials were supplemented by some ‘armchair fieldwork’ among the Angolan and Congolese diaspora in Belgium, whereby the basic-vocabulary equivalents for the southern Kisikongo (as spoken in Mbanza Kongo, for which there was no recent data) and the N’zeto variety, as well as the central Kindibu (missing from Bastin et al. (1999)) were collected. A further 81 ‘doculects’ were selected, with a ‘documented lect’ or doculect meaning “a linguistic variety as it is documented in a given resource” (Cysouw & Good 2013:342). Those doculects include five languages belonging to zone A plus B25 Kota, eight languages belonging to zone C, and a selection of B40-B50-B60-B70-B80, H10-H20-H30-H40, K10-K30, L10-L50 and R10-R20-R30 languages. All Bantu varieties chosen are thus North-West and West Bantu.

Altogether, this resulted in a dataset of 95 Bantu varieties, with for each of the 92 basic-vocabulary concepts one or more equivalents (only occasionally is an equivalent missing). See the online Supplementary material for the details. With regard to the 82 doculects, it needs to be pointed out that only the raw, underlying data was used, not anyone else’s cognition judgements. In effect, all the cognition judgements—i.e., the identifications of the number of cognates for each of the 92 concepts—were done from scratch, starting from the 2012 and 2013 fieldwork data, and progressing towards the neighbouring varieties. In reusing the Tervuren data, for instance, this thus meant that up to 92 basic-vocabulary items (rather than a maximum of 90; cf. Section 1.5.1.3) were considered. The Tervuren lists were also ‘unlumped’ (i.e., the collation of different doculects was undone, and one doculect only was chosen per variety).

23. Our Cimbala may actually be (or be related to) van Bulck’s (1948:362) ki(mi)Mbala, his number 158 (cf. Section 1.2.1, Figure 6).
24. The raw Tervuren data may be found at http://www.africamuseum.be/research/
To account for lexical variation we predominantly proceeded as is customary in lexicostatistic studies, that is, on the level of cognates or words having a common etymological origin. Most Kikongo varieties, for instance, have a common term for the concept ‘to kill’, as shown in (1a) with KongoKing fieldwork data. The consonantal variation observed in stem-initial position is the outcome of the regular sound change known as *p*-lenition (Bostoen et al. 2013:66). The different verb stems in (1a) are therefore considered to be cognate and are also encoded in the same way. They could be reflexes of the etymon reconstructed as *-pón*- ‘pound, stamp, beat; forge’ in BLR3 (Bastin et al. 2002), though on the basis of reflexes attested in the East Bantu zones E, G, L, J, M, N, P and S. Some Kikongo varieties, such as Kibembe in (1b) and the B40 varieties in (1c), do have clearly distinct and unrelated translation equivalents for this basic concept, which are consequently encoded differently.

(1) a. Kisikongo -vonda (KongoKing fieldwork 2013) 
   Kisolongo -vonda (KongoKing fieldwork 2012) 
   Kindibu -vonda (KongoKing fieldwork 2013) 
   Ciwoyo -vonda (KongoKing fieldwork 2012) 
   Cizali -vonda (KongoKing fieldwork 2012) 
   Cimbala -vonda (KongoKing fieldwork 2012) 
   Kimbata -yonda (KongoKing fieldwork 2012) 
   Kimbeko -yonda (KongoKing fieldwork 2012) 
   Kinkanu -yonda (KongoKing fieldwork 2012) 
   Kimpangu -yonda (KongoKing fieldwork 2012) 
   b. Kibembe -júsà (Jacquot 1981) 
   c. Punu -boka (Mavoungou & Plumel 2010) 
   Lumbu -boka (Mavoungou & Plumel 2010)

In order to exploit lexical variation maximally for the establishment of internal sub-grouping, we have sometimes proceeded to a more fine-grained differentiation than would usually be done in a Bantu-wide study of the same kind.25 Certain lexical items that can ultimately be considered cognate on a deeper Bantu level have been judged as historically distinct in our study along two parameters, namely derivation and irregular sound change. An illustrative case in which verbal derivation was taken into account to diversify the innovations maximally is the concept ‘to give’. As shown in (2a), the Proto-Bantu base verb *-pá- has only a few retentions in Kikongo. Most varieties replaced it by a reflex of *-páan- derived from *-pá-, as shown in (2b). Although associative/reciprocal in origin, reflexes of *-páan- no longer have a reciprocal meaning in Kikongo and also behave syntactically in the same way as reflexes of the ditransitive source verb *-pá-. The derivational semantics and intransitive argument structure went lost. The reflexes of *-páan-functioning as basic ‘give’ verb can therefore be encoded differently from the direct reflexes of *-pá-. Likewise, some Kikongo varieties have a third ‘give’ verb which is possibly still another derivation from *-pá-, namely the non-reconstructed *-pá-ik-.

human-sciences/cultsoc/lexico-1/.

25. Compare in this regard Ehret & Kinsman (1981:402) whose approach to cognation judgement was “carried out with an eye to bringing dialect differences into sharp relief”.
resulting in reflexes of an underlying form */-péek-*, as illustrated in (2c). This might originally have been a ‘neuter’ derivation of */-pá-* meaning ‘to be in a state of giving’, which (in analogy with the associative/reciprocal */-páan-*) eventually acquired the active meaning and ditransitivity of its original base verb. In spite of this common etymological origin, it can therefore also be encoded differently.

(2) a. Kibembe -ha (Jacquot 1981)
    Kidondo -ha (Mfoutou 1985)
    Kizombo -vá (Bastin et al. 1999)

b. Kisikongo -váána (KongoKing fieldwork 2013)
    Kindibu -vaána (KongoKing fieldwork 2013)
    Ciwoyo -vaana (KongoKing fieldwork 2012)
    Cimbala -vaana (KongoKing fieldwork 2012)
    Kimbata -yaana (KongoKing fieldwork 2012)
    Kimboko -yaana (KongoKing fieldwork 2012)
    Kinkanu -yaana (KongoKing fieldwork 2012)
    Kisundi -gana (Bastin et al. 1999)
    Punu -vhaana (Mavoungou & Plumel 2010)
    Lumbu -vhaana (Mavoungou & Plumel 2010)

c. Kisundi -hééka (Bastin et al. 1999)
    Punu -vheegha (Mavoungou & Plumel 2010)
    Lumbu -vheegha (Mavoungou & Plumel 2010)

If sound change is regular, it has not been taken into account to quantify lexical variation. However, if alterations in the shape of lexical items are at odds with regular sound correspondences, they have been considered as lexical innovations and were encoded differently. The Kikongo lexical items found for ‘egg’ constitute a good case in point here. In certain varieties, as shown in (3a), a direct reflex is found of */gɪ́*, which is considered to be a ‘main reconstruction’ in BLR3. More Kikongo varieties, as those in (3b), have a reflex of the form reconstructed as either */jágɪ́* or */jákɪ́* in BLR3. The variation in C2 points to the fact that present-day /k/ can be a regular reflex of both */k/ and */g/. Bastin et al. (2002) consider */jágɪ́* and */jákɪ́* as ‘VARiants’, thus reconstructions descending from the more basic, main etymon */gɪ́*. However, since this can impossibly be done in a phonologically regular way, we have encoded them as two distinct lexical items. Forms such as those in (3c), all ending in /o/, for which no reconstruction was proposed in BLR3, are possibly also historically related to */gɪ́*, but certainly not in a phonological regular way and are therefore also encoded differently.

(3) a. Kimanyanga diiki/maaki (Makokila Nanzanza 2012)
    Cilinji diiki/maaki (KongoKing fieldwork 2012)
    Kizobe diiki/maaki (KongoKing fieldwork 2012)
    Kiyombe dííki (De Grauwe 2009)
    Kibembe dí:̀kì/mí:̀kì (Jacquot 1981)
    Cilaadi li’ki/mè’kì (Jacquot 1971)
    Lumbu diki (Mavoungou & Plumel 2010)

b. Kisikongo dyáki/mááki (KongoKing fieldwork 2013)
    Kindibu diaki/maaki (KongoKing fieldwork 2013)
Similarly for the concept ‘head’, while all lexical items are ultimately reflexes of *-tʊ́è, a distinction was made between those having a CV stem vs. those with a CGV stem, mostly ending in a front vowel (as in Kindibu ntu/zintu vs. Kinkanu ntwa). Or for ‘hear’, where most if not all are eventually reflexes of *-jígu, a distinction was made between those where the stem has become a simple glide and those where more is maintained (as in Cimbala -wa vs. Cizali -yuwa).

In contrast to traditional lexicostatistic studies, any number of cognates per concept was also allowed. As can be seen in (3), for instance, Cilinji and Kizobe have two distinct noun stems for ‘egg’, both of which were taken into account. In the event, up to four cognates were recorded for some concepts. Across the 92 concepts, and for the 95 Bantu varieties, a total of 1,324 cognate sets, coded as discrete binary characters, were eventually obtained.

2.2. Methods

We have used two phylogenetic methods, to first produce a network, followed by a tree, in accordance with: “In practice, network-type methods have often been used just as initial diagnostic tests, to judge whether a particular dataset yields a pattern that is tree-like ‘enough’ in order to justify proceeding to tree-only analyses useful for particular research ends” (Heggarty et al. 2010:3835). In our case, we have opted for the distance-based Neighbour-Net to draw our network, and for the character-based Bayesian MCMC to draw our tree. Brief definitions of these methods have been presented in Section 1.5.1.1; their position within the range of phylogenetic approaches to language classification is as shown in Figure 23.
The Neighbour-Net network was straightforwardly produced using SplitsTree4 (Huson & Bryant 2006), which is software for computing phylogenetic networks. The Bayesian tree representation was produced using a Markov chain Monte Carlo approach (Larget & Simon 1999). The phylogenetic inference was performed using BayesPhylogenies (Pagel & Meade 2004) with a covarion model (Tuffley & Steel 1998, Penny et al. 2001) applied to the data, which allows for jumps in the rate of evolution in different sections of the tree. Such a model follows the belief that languages do not change at a constant rate (McMahon & McMahon 2005) which is supported by empirical work that has examined how rates of change vary (Pagel et al. 2007, Atkinson et al. 2008, Greenhill et al. 2010). The covarion model allows the rate of change to switch between different categories of rate; for instance, a two-state covarion model will estimate a slow and a fast rate category and also how fast the model switches between these two categories. The n-state covarion model expands this to estimate n rate categories and the switching rates between every pair of these categories. Model testing gives us a procedure to test whether the addition of these extra parameters into the model of evolution produces a significant improvement to the model. We used a stepping stone sampler (Xie et al. 2011) to perform the testing and this showed the four-state model to be the best. See the Supplementary material for the details.

An exponential branch length prior was used (mean = 0.1). Testing showed that the choice of mean for the exponential prior did not change the topology, nor did using a uniform prior. Also, as per the recommendations of Brown et al. (2010) BayesPhylogenies uses suitable starting branch lengths and a whole-tree scaling function.

We rooted our tree with North-West Bantu languages, viz. A44 Tunen, A75a Fang, A81 Kwasio, A86c Mpiemo, A91 Kwakum and B25 Kota. Five experiments were each run for 80 million iterations, with an initial burn-in period of 30 million iterations and post run analysis of the parameters and log-likelihoods revealed that all experiments had converged. Samples were taken at 10,000 iteration steps to avoid auto-correlation. The final tree was produced from 1,000 samples taken from the end of the chain. The experiments were each run on one node consisting of two six-core Intel processors. Each experiment took approximately eight hours.

3. Our phylogenetic classification: Results

3.1. Neighbour-Net

The Neighbour-Net network of all 95 West Bantu varieties used in our study is shown in Figure 24. Splits clearly set apart (i) the zone A and C languages plus B25 Kota, (ii) the zone B languages minus B40, (iii) the cluster of Kikongo varieties which are seamlessly integrated with B40 languages and more distantly with some L10-H30-H40 languages, and (iv) the remainder (L50-H20-K-R). Important here is that the Bantu varieties in (iii) cluster together. The internal connections within (iii) may be studied further in the extract shown in Figure 25.
Figure 24. Neighbour-Net network of the 95 West Bantu varieties used in our study.
Figure 25. Neighbour-Net network of the Kikongo and related varieties used in our study.
What was already apparent from Figure 24 is even more so in Figure 25, where colours depicting the geography on the ground were added for ease of interpretation. The most marked sub-group is the one of the western Kikongo varieties, for which there are only a limited number of optional paths (technically: ‘reticulations’) at the root. Also set apart, with more optional paths at the root, are the Shira-Punu languages on the one hand, and the group of Kiyaka, Kisuku, Hungan and Saamba on the other. The southern, central and eastern Kikongo varieties form another sub-group, with as outliers the two Kisolongo varieties on the one hand and the sub-group of eastern varieties on the other. The last sub-group brings together the northern Kikongo varieties, where Cilaadi and a neighbouring Kisundi variety occupy an intermediate position between the other northern varieties and the sub-group of eastern varieties. On the face of it, the picture that emanates from Figure 25 makes a lot of sense. If the goal of a network is to show that language varieties typically don’t have single parents and are instead the result of multiple connections between various varieties, then the internal sub-groups seen in this network correspond well with the geography on the ground. In doing so one must also recall that a two-dimensional network, where varieties are organised in circular fashion, can only place any one variety between two other varieties. To go from a western Kikongo variety to one from the B40 sub-group, for example, one must therefore realize that one can also walk through the network, following up the links inside the mesh of optional paths. In any case, and following Heggarty et al. (2010:3837), we are tempted to assume (although this remains undemonstrated) that the Neighbour-Net shown in Figure 25 has managed to tease apart tree-like vs. web-like components, with the splits suggesting migrations, the waves suggesting dialect continua.

These results indicated that an attempt to draw up a tree using advanced computational tools and ditto computer power was thus a worthwhile venture.

3.2. Bayesian MCMC

The Bayesian consensus tree for our 95 West Bantu varieties is shown in Figure 26. The numbers shown on the tree are the Bayesian posterior probabilities, which indicate how often each of these nodes was observed in the sample. We have not shown the percentages of the nodes near the tips. The branch lengths are in lexical change, so a longer branch indicates that more change has been observed in that part of the tree.

As may be seen, our tree corresponds very well with our network. The tree-like signals that were suggested in the network reappear as separate clades in the tree. The Kikongo and related varieties, which have been coloured in Figure 26, clearly form a cluster, which is why we will henceforth refer to this clade as the Kikongo Language Cluster, for short KLC.
Figure 26. Bayesian consensus tree of the 95 West Bantu varieties used in our study.
4. Our phylogenetic classification: Discussion

4.1. External classification: the KLC vs. neighbouring clades

The languages from Zone A, plus B25 Kota, were used to root our tree. These cluster together at the top. Zone C languages, spoken in the Inner Congo Basin, branch off next. Then follows the so-called West-Coastal group: a number of clusters with zone B languages and the KLC. The last cluster in our tree is known as South-West Bantu, here with representatives from groups H20-H30-H40, L10-L50, K10-K30, R10-R20-R30. The terms Inner Congo Basin, West-Coastal and South-West are inspired by Vansina’s (1995) interpretation of the various lexicostatistic trees found in Bastin et al. (1999), to which he had advance access. Together, these three groups are called “narrow West Bantu, i.e., West Bantu minus the North-west” by Vansina (1995:186). What our tree shows (and confirms, cf. Vansina (1995) and Bastin et al. (1999)) is that Kikongo (and by extension the KLC) forms part of West-Coastal, and that the immediate external neighbours of West-Coastal are the Inner Congo Basin on the one hand, and South-West on the other.26 What this external classification also indicates is that Kikongo’s nearest southern neighbour, namely Kimbundu, is thus phylogenetically quite distant from the KLC.

In terms of interpreting our Bayesian tree, it must be pointed out that after its generation this tree was also “ladderised”. This standard procedure means that while all splits are about two groups separating, whether they are shown on the top/bottom (or left/right) of this split is mostly a visual choice. What is grouped together, how close groups are to each other (i.e., how many other branchings there are in-between groups), and the branch lengths represent how closely related varieties are. A ladderised tree is therefore not truly directional, with the x-axis reflecting a relative time axis (Omland et al. 2008:863). Care should thus be taken not to blindly read the sequence of the clades as the passage of time. Our data, then, is telling us that we have a style of tree where small groups (or single taxa) have split off from the main backbone of the tree, instead of a tree where we have repeated big groups splitting into two equal sub-groups.

Specifically for West-Coastal, our tree shows three distinct clades: (i) B50-B60-B70 languages, (ii) B80 languages, and (iii) the KLC. These three clades have been mapped in Figure 27. Using the ‘centre of gravity’ model (Wichmann et al. 2010), one can state that the greatest linguistic diversity within the West-Coastal Bantu group is located somewhere to the north-east of the Pool region, as it is there that the three clades seen in Figure 27 come together. Regarding the location of the homeland of West-Coastal Bantu, it has been suggested that:

26. Although not of immediate relevance to our study, and not in contradiction with what was just said, it must be pointed out that in an ongoing Bantu-wide study (Grollemund et al. 2015), which includes a very large selection of our Kikongo varieties, South-West Bantu together with all East Bantu languages is coordinate with the whole of West-Coastal (and not merely with the KLC as in Figure 26). This view can only be obtained by adding East Bantu languages. The fact that the addition of East Bantu results in the flotation of South-West, likely signals the long historical contact between the savannah languages—i.e., between South-West and languages from the centre-east. As Bastin et al. (1999:112) correctly pointed out: “There is always a risk of distortion at the cut-off made in a continuum.”
“The fragmentation center of this subclade [i.e., West-Coastal Bantu] is difficult to pin down with exactitude, but it was probably situated somewhere in between the Batéké Plateau and the Bandundu region, that is, around 3°S and between about 14°E and 17°E” (Bostoen et al. 2015:361). Given that the three clades of West-Coastal come together there, and not in the Lower Congo region where one finds the KLC, one can state that the Lower Congo region is an area of later settlement for this subset of West-Coastal Bantu languages rather than a West-Coastal homeland. Regarding the issue of whether or not there is such a thing as a ‘Kongo-Nukleus’ (Heine 1973, Heine et al. 1977), as discussed in the introduction, we can now state that the Lower Congo region was not a centre of expansion. Based on this new evidence, we can also claim that the direction of migration was from the hinterland towards the Lower Congo region, not from the Coast towards the hinterland.

Figure 27. West-Coastal Bantu with its three distinct clades in our Bayesian consensus tree.

4.2. Internal classification: sub-groups within the KLC

The Kikongo (KK) Language Cluster, which constitutes a discrete clade within West-Coastal Bantu, may be subdivided into four distinct sub-groups, as shown in Figure 28, labelled geographically as North KK, East KK, South KK and West KK. The varieties within each of these sub-groups all happen to be geographic neighbours of one another. The Bayesian posterior probability at the node from which the centrally-located Kindibu and Kimanyanga branch off is the lowest of the entire KLC, however: in only 68 out of every 100 trees in the sample do these two
languages form a coordinate branch with the other South KK varieties. In effect, from the network seen in Figure 25, these two varieties are ‘as close’ to South KK as they are to East KK, and walking through the mesh of the network, they are even closer to the northern Cilaadi and a neighbouring Kisundi variety than they are to the two southern Kisolongo varieties. We therefore wish to view the area of the two centrally-located Kindibu and Kimanyanga varieties as a buffer zone between the South KK, East KK and North KK sub-groups. The status of this Central KK sub-group is comparable to what Bastin & Piron (1999) called a groupe flottant ‘floating group’—in their case a group which moves about in various trees, depending on the type of tree calculation used.

**Figure 28.** The Kikongo Language Cluster (KLC), with its Kikongo sub-groups highlighted.

As far as the core of the KLC is concerned, Figure 28 indicates that North KK forms a coordinate branch with all other Kikongo varieties. These other varieties split into East KK and South KK on the one hand, and West KK on the other. Note, however, that the posterior probabilities of nodes 3 and 4 are also pretty low, at just 72. This suggests that although the KLC phylogenetic sub-groups themselves are solid, the way they relate to each other is less certain and needs to be tested against other types of linguistic evidence, such as phonological and morphological innovations (Dom & Bostoen this issue). Figure 28 may be represented schematically as shown in Figure 29.
Figure 29. Schematic representation of the relationship between the sub-groups within the KLC.

Viewed in this way, one thus recognises a ‘backbone’ of migrating people—the numbers 1, 2, 3 and 4 in Figure 29, which correspond to the same numbers at the nodes in Figure 28. What is moreover crystal-clear from our results is that the B40 languages, spoken along the coastal regions of Congo-Brazzaville and especially Gabon, are an integral part of the KLC. They are fully connected to the Civili and Kiyombi as spoken in Congo-Brazzaville, for which the Bayesian posterior probability at the connecting node is 100—node number 5 in Figure 28 and Figure 29. All of these varieties together are also well connected to the rest of West KK, for which the Bayesian posterior probability at the higher connecting node is an impressive 86. While the network in Figure 25 set apart the B40 languages from the other West KK languages (albeit that they are actually close to one another when walking through the mesh), Figure 28 indicates that the B40 languages thus (always!) go together with Kiyombi and the Civili varieties. We have chosen to term this last sub-sub-group West KK Ctd. in Figure 29.

At the other end of the KLC spectrum, a number of languages branch off first, before the four distinct KK sub-groups appear. These languages are part and parcel of the KLC, and we henceforth wish to view these languages—in our sample L12a Saamba, H42 Hungan, H31 Kiyaka and H32 Kisuku—as Kikongoid languages: they are phylogenetically close, but are not part of ‘core Kikongo’.

The first result, namely that the B40 languages from Congo-Brazzaville and Gabon may be closely related to zone H or even group H10 languages, has been suggested a number of times in the past, starting with Achille Meeussen (1980:604), to Bastin & Piron (1999:159-60), and most recently Grollemund (2012:247). Linguists from Congo-Brazzaville such as Bouka & Ndamba (1991) also noted the affinity (see Section 1.4), while the makers of the Atlas linguistique du Congo, which covers both Congo-Brazzaville and the present DRC, pointed out for B44 Lumbu and B45 Bwisi (see Figure 16): “cette classification (en zone B), est contestée par les locuteurs-natifs et est contradictoire avec leurs traditions orales. L’appartenance au groupe koongo (H.10) est revendiquée” (ALAC 1987:38). In a study based on the intercomprehension of the Gabonese Bantu languages, Kwenzi-Mikala (1988) brings together H12 Civili and the B40 languages into a
sub-group which he calls ‘Merye’. In hindsight then, the century-old classification of Johnston (1922) was surprisingly spot-on in this respect, as his Kongo group saw B44 Lumbu as a natural extension of H12 Civili (cf. Figure 4). Bastin & Piron (1999:160) moreover wondered: “les affinités qui apparaissent dans la classification entre les parlers classés en B.40 et la zone H reflètent-elles un lien généalogique ou sont-elles liées aux aires de contact?” Insofar as lexically-driven phylogenetic language classifications may attempt to provide an answer to this question, a more detailed study of the KLC network in Figure 25 is in order. This network shows that while there are indeed quite some optional paths within the H10 sub-group of West KK, as there are within the B40 sub-group of West KK, both these sub-groups have hardly any parsimonious optional paths connecting them through the mesh. This, then, indeed points towards a genealogical link rather than being the result of language contact. The language varieties within the H10 sub-group of West KK may be seen as a continuum of varieties, as can those within the B40 sub-group of West KK, but the historical link between these two sub-groups is the result of (a huge) migration. Figure 28 further suggests that the West KK sub-group actually consists of (i) Cimbala and Cizali—two riparian varieties, on the right bank of the mouth of the Congo River, across from and adjacent to South KK varieties—who branch off first, (ii) H10 varieties in the DRC and Cabinda, and (iii) H10 and B40 varieties in Congo-Brazzaville and Gabon.

The second result, that several languages from the Kwilu-Kwango region are closely related to the core KLC and may be termed Kikongoid, is a surprising finding. All said, it is true that both Hulstaert (1950) and van Buleck (1954) included languages like Kiyaka and Kisuku among their Kikongo language varieties, but then their Kikongo groups were so encompassing that they also included languages which decidedly belong to other clades, in casu South-West Bantu, such as H21a Kimbundu, L11 Pende and L12b Holu in Hulstaert (1950), or L13 Kwezo (Pindi) in van Buleck (1954). Languages like Kiyaka, Kisuku, Hungan and Saamba, although outliers within the KLC, are thus more related to Kikongo than they are to the neighbouring B80 group of languages to their north-east. The exact implication of this could now constitute the starting point for further research. This is beyond the intended scope of the present article, however.

The most novel result of our study is the actual internal classification of the core KLC, which was never before attempted on this micro-scale, except for Ntunda Nzeza’s (2007) effort limited to the DRC. While it is true that Laman (1936), who worked chiefly on Kimanyanga, looked around from his vantage point and ended up labelling his own Kikongo variety as the central one and termed some others with cardinal directions from there (cf. Figure 5), this was simply a fortuitous result of the location of the Mukimbungu mission of the Swedish Missionary Society from where he operated. Similarly, the sub-groupings seen in van Buleck (1954) are merely attempts to group the material geographically. Instead, our KLC sub-groups emerged without any a priori vantage point; given the outcome, they were labelled with the cardinal directions. In order for us to now interpret our results further, Figure 30 maps the sub-groups that we uncovered.

27. This sub-grouping is however contested by several other Gabonese linguists (Mavoungou & Ndinga-Koumba-Binza 2010:8).
Even though one may have wanted even more sampling points, it is immediately clear that rivers played an important role in the internal diversification of the KLC, mainly the Congo and Inkisi Rivers. East of the Inkisi River one finds all the Kikongo varieties from the East KK sub-group; north of the Congo River one finds all the Kikongo varieties from the North KK and West KK sub-groups; and south of the Congo River one finds all the Kikongo varieties from the South KK sub-group. The only Kikongo variety which does not follow this geography is DRC Kisolongo, spoken by the riparian Basolongo on the right bank near the mouth of the Congo River. Together with Angolan Kisolongo—its sister variety spoken right across the Congo River, near Mbanza Soyo and further inland—it belongs to the South KK sub-group. Despite this joint sub-grouping, do observe the large difference in branch length between these two Kisolongo varieties, as may be seen from Figure 28, which indicates some serious lexical change between those two varieties nonetheless. This may be interpreted that while they were once quite close (and very likely one and the same language variety, of course) they are rapidly busy diverging. Another river that may have played a role is the Shiloango River, dividing the southern West KK varieties in our sample from the northern West KK varieties.

---

28. For instance, on the map shown in Figure 1, one also notices the Kikongo varieties Kikamba, Kitsootso and Dihungu, but (reliable) basic vocabulary could not be collected for them. These varieties have nonetheless been included on the map with the contribution of Dom & Bostoen (this issue) in mind.
We are now in a position to evaluate the classification proposed by Ntunda Nzeza (2007). On the map shown in Figure 17, we coloured Ntunda Nzeza’s language codes in such a way that they reflect our phylogenetic classification. Those same colours were then added to his classification shown in Figure 18. As may be seen from the intertwining of the colours, none of our phylogenetic sub-groups surfaces in his classification. In an attempt to map the classifications onto one another, one could allow any of the Central KK varieties to go with a neighbouring sub-group, at which point Ntunda Nzeza’s A2β corresponds with our North KK. His B1β and B2β would then also constitute parts of East KK. However, beyond these efforts, it is clear that there can be very little validity in Ntunda Nzeza’s classification. His A1α even brings together varieties from as many as three sub-groups, spoken at the extreme ends of the KLC in the Lower Congo province of the DRC: from Kisolongo (SO) in the west, over Kiwoyo (WO) and Kizombo (ZO), to Kiluula (LU3) in the east. This is highly unlikely, as pointed out by Masele & Nurse (2003:122): “High levels of lexical similarity go together with physical proximity, low figures correlate with distance, a conclusion reached by many other observers in Africa” (see also Henrici 1973:90, Nurse 1997:365).

What can our internal division of the KLC now tell us about Kongo history? According to 16th century oral traditions, the origins of the Kongo kingdom can be traced back as far as the second half of the 14th century, beginning around 1350-1375 (Thornton 2001:119). A phylogenetic classification that takes as its input the basic vocabulary is supposed to be able to go back much further in time, revealing deep historical breaks. Indeed, the fragmentation of the core KLC is a phenomenon that considerably pre-dates the birth of the Kongo kingdom. A preliminary attempt to date the fragmentation of the KLC suggests that the four major sub-groups—i.e., North KK, East KK, South KK and West KK—were already fully in place 1,500 years ago (cf. de Schryver et al. 2014, where a maximum clade credibility tree was calibrated using historical Kikongo sources on the one hand, and archaeological evidence on the other). We know from the historical sources (Thornton 2001) that the rise and apogee of the Kongo kingdom took place within the confines of the South KK sub-group, in close association with polities situated within the confines of the East KK sub-group (i.e., east of the Inkisi River), which later became the kingdom’s Mbata province. While we have documented (i.e., ‘written’) evidence for the start of the spread of certain phonological features in the 17th century, from the heart of the Kongo kingdom in Mbanza Kongo in the south towards regions immediately north, east, and west of it (Bostoen & de Schryver 2015)—towards respectively Central KK, East KK and West KK—proto East KK and proto West KK had already been diverging into new varieties for about a thousand years (de Schryver et al. 2014). The Kikongo variety spoken at the capital Mbanza Kongo was thus not exported towards all the regions of the expanding kingdom. Rather, some (prestige) features did spread, but only modified core KLC varieties that were already being spoken in those regions. In actual fact, several other closely related kingdoms, such as Ngoyo, Kakongo and Loango, arose within the confines of the West KK sub-group (Proyart 1776); see Figure 2 (with the Shiloango River ‘Louango Louiza’ separating the southern Ngoyo and Kakongo kingdoms from the northern Loango kingdom).
The North KK sub-group was not home to centralised polities, but remained—from the viewpoint of the surrounding kingdoms—‘in the periphery’ and also constituted a contact zone with the B70 Teke varieties. However, the Kongo kingdom’s northernmost province Nsundi, which also started out as an independent polity, considerably expanded to the northeast along the south banks of the Congo and across during the late 15th and early 16th centuries (Hilton 1985:55, Thornton 2001:115), possibly into territories that used to be populated by speakers of North KK varieties.

The Central KK buffer zone, finally, indicates more recent contact, between South KK varieties and East KK varieties, between East KK varieties and North KK varieties, and even between South KK varieties and North KK varieties. The emergence of that central contact zone may tentatively be dated to about 500 years ago (de Schryver et al. 2014), which means that here the effects of the expanding Kongo kingdom started to have an impact on the languages of that region. These central varieties do not form their own separate clade (yet), however. That said, it has been shown that for borrowings to have a phylogenetically measurable effect, one needs a great amount of horizontal transmission (Greenhill et al. 2009), so the fact that we do notice the unstable behaviour of these two varieties does indicate that there has been a considerable amount of contact and borrowing in that region. Reformulated, over the past 500 years the ancestor varieties of today’s Kimanyanga and Kindibu ‘caught’ enough innovative basic vocabulary from varieties in the neighbouring sub-groups so as to behave in an unstable way in the phylogenetic tree. It is tempting to see Kimanyanga as an initially northern variety and Kindibu as an initially southern one.

Given the deep historical break-up of the KLC into four sub-groups, a process that preceded the emergence of the Kongo kingdom itself by a wide margin, one may of course rightfully wonder why the KLC should be called the KLC at all. Viewed from the proto KLC angle, a parent language to the core KLC as well as the Kikongoid languages—number ① in Figure 28—may be dated to about 3,000 to 2,000 years ago (de Schryver et al. 2014). There is nothing Kikongo about that proto language, and definitely nothing relating it to a Kongo kingdom. Conversely, that proto language did eventually lead to a huge cluster of languages that we now call Kikongo, as well as closely related languages like Kiyaka, Kisuku, Hungan and Saamba. Therefore, while the term KLC is surely Kongo-centric, it is at least a term which also implicitly refers to the most illustrious of all the pre-colonial Central African polities, known to the West since 1482 and raising more political, commercial, religious and academic interest than any other.

To conclude this discussion, we would like to offer tentative but parsimonious migration routes which explain the emergence of the various sub-groups which we uncovered for the KLC. This is done in Figure 31, using the backbone numbers from Figure 28, and thus taking for the time being the hierarchy between the phylogenetic sub-groups depicted there as the actual phylogeny. Starting about 3,000 to 2,000 years ago, a Kikongoid branch broke off from proto Kikongoid somewhere due north of the Pool area ①, with its speakers going around the Pool, southwards. This eventually led to the emergence of Kikongoid languages like Kiyaka, Kisuku, Hungan and Saamba. Following the Congo River on its north side,
North KK split off from proto Kikongo around the area where Cilaadi is currently spoken ②. In due time, this gave rise to northern varieties like Kisundi (in Boko), Kihangala, Kibembe, Kidondo, Kikunyi and Kisundi (in Kimongo). The most westward of these may have been in contact with West KK varieties at some later point. Meanwhile, speakers of Kikongo languages continued moving southwards, and around our Kimanyanga sampling point, the West KK branch broke off ③, continuing their migration south-westward on the north side of the Congo River, all the way to the mouth of that river and then up north again. This gave rise to varieties like Cimbala, Cizali, Ciwoyo, Iwoyo, Cisundi, Kiyombe, Kizobe and Cilinji. Others effectively crossed the Congo River around our Kimanyanga sampling point and continued their south-westward move on the south side of the Congo River. Hitting the Kwilu River ④, some branched off to the east and crossed the Inkisi River, which led to East KK varieties like Kintandu, Kimpangu, Kimbata, Kimbeko and Kinkanu, while others migrated on the south side of the Congo River, which led to South KK varieties like Kiboma, Kisolongo, Kisikongo, Kizombo, etc. In the west, the West KK continued their migration along the Atlantic coast, crossing the Shiloango River ⑤, giving rise to Civili varieties, Kiyombi and the various B40 languages. As argued above, during the rise of the Kongo kingdom, and as a result of its expansion, the central contact zone emerged, with Central KK varieties like Kindibu and Kimanyanga.

Figure 31. Postulated migration routes based on the internal classification of the KLC.
Conclusions

In this article we set out to show how more than half a century of vocabulary-based and vocabulary-driven classifications for the Bantu languages led from a heavy use of lexicostatistic methods to the phylogenetic approaches of today. The various lexicostatistic trees obtained by the Tervuren linguists, in close collaboration with computational help from London, were by far the most advanced of their era. Their Bantu-wide variable composite heterogloss maps and their Variable Neighbourliness trees produced at the close of the millennium (Bastin et al. 1999) were the absolute highpoint of Bantu lexicostatistics, and will very likely never be bested within the traditional lexicostatistic framework. Starting early in the new millennium, the first vocabulary-driven phylogenetic classifications began to appear for the Bantu languages (Holden 2002, Holden & Mace 2003, 2005, Holden et al. 2005, Holden & Gray 2006, Pagel & Meade 2006, Walker & Hamilton 2010, Dunn et al. 2011, Opie et al. 2014), with also an attempt to include grammatical features into the model (Rexová et al. 2006) as well as a recent phylogeographical attempt (Currie et al. 2013). All of these phylogenetic classifications for the Bantu languages at large rather blindly reused both the data and the cognition judgements from Tervuren. Grollemund (2012) was the first phylogenetic study for (North-West) Bantu which was based on a new lexical data set, for which fresh cognition judgements were made. Our present study continues this line, in that novel data coupled with fine-mazed cognition judgements were assembled, in order to draw up a tree for narrow West Bantu with a particular focus on the Kikongo (KK) varieties spoken in the Lower Congo region and beyond.

Our study—in which Neighbour-Net was used to draw up a network, and a Bayesian inference of phylogeny with a Markov chain Monte Carlo approximation algorithm for the tree—has revealed the existence of what we termed the ‘Kikongo Language Cluster’ (KLC). The KLC is a disparate continuum of closely related Bantu languages spoken in the wider Lower Congo region and constitutes a discrete clade within West-Coastal Bantu. In addition to the H10 languages, this clade fully encompasses the B40 languages spoken as far away as Gabon, and even includes what we termed ‘Kikongoid’ languages like Kiyaka, Kisuku, Hungan and Saamba spoken in the Kwilu-Kwango (i.e., some H30 and H40 languages, but not H20 languages). The core KLC further fragments into four neat subgroups, for which we use cardinal labels: North KK, East KK, South KK and West KK. The Congo and Inkisi Rivers agreeably divide the Kikongo varieties within these sub-groups, with North KK and West KK strictly to the north of the Congo River, East KK strictly to the east of the Inkisi River, and South KK strictly to the south of the Congo River (bar one riparian variety). Parent languages of each of these four sub-groups and the sub-groups themselves were already in existence about 1,500 years ago. This time-depth thus considerably predates the emergence of the Kongo kingdom, which oral traditions place in the second half of the 14th century. The state-formation processes that accompanied the rise of the Kongo kingdom have thus no bearing on the existence of those four sub-groups within the KLC, nor have the consequences of the Civil War during the second half of the 17th-early 18th century (Thornton 1983).
There is one region where the expanding Kongo kingdom did have a noticeable classificatory effect on the internal divisions of the KLC, however, as it even led to the emergence of a contact zone where South KK, East KK and North KK met, creating a convergence zone. We termed this emerging contact zone, situated north and south of the unnavigable stretches of the Congo River, Central KK.

Finally, our phylogenetic classification of the Kikongo Language Cluster has also allowed us to propose the tentative migrations which underlay its emergence. Through it, we have begun to rewrite the early history of the wider Lower Congo region, and it is now hoped that our classification can form the basis for subsequent studies.

Acknowledgements

The research for this article has been funded by the European Research Council through a Starting Grant awarded to KB (KongoKing, Grant No. 284126), as well as the Special Research Fund of Ghent University for KB and G-MdS. Mark Pagel’s Advanced Investigator Award from the European Research Council (Mother Tongue, Grant No. 268744) supported RG, SB, and the advanced computational processing. We would like to thank Sebastian Dom and Heidi Goes, as well as the two anonymous referees, for comments on an earlier version of this article. Thanks are also due to Rik Declercq for his help with the digitization of the larger maps reproduced in this article.

References


Ekholm, K. 1972. *Power and Prestige, the rise and fall of the Kongo kingdom*. Uppsala: Akademisk avhandling, Skriv Service AB.


Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C. & Pagel, M. 2015. ‘Bantu expansion shows that habitat alters the route and pace of human


Hymes, D.H. 1960. ‘Lexicostatistics so far’. *Current Anthropology* 1,1: 3-44.


Laman, K.E. 1936. *Dictionnaire kikongo-français, avec une étude phonétique descrivant les dialectes les plus importants de la langue dite kikongo* (Institut royal colonial belge, Section des sciences morales et politiques, Mémoires - Collection in-8°, Tome II). Brussels: Georges van Campenhout.


Mann, M. 1970. ‘Internal relationships of the Bantu languages: Prospects for topological research’. In D. Dalby (ed.), *Language and History in Africa: A volume of collected papers presented to the London seminar on language and


Tunks, J. 1993. ‘A Lexicostatistical Study of Southern Bantu Languages’. Department of Linguistics, University of California at Los Angeles, Los Angeles, CA.


**Authors’ addresses**

Gilles-Maurice de Schryver  
KongoKing Research Group  
Department of Languages and Cultures: Africa  
Ghent University  
9000 Ghent, Belgium  
Department of African Languages  
University of Pretoria  
Pretoria 0002, South Africa  
gillesmaurice.deschryver@UGent.be

Rebecca Grollemund  
Evolutionary Biology Group  
School of Biological Sciences  
University of Reading  
r.b.grollemund@reading.ac.uk

Simon Branford  
Evolutionary Biology Group  
School of Biological Sciences  
University of Reading  
s.j.branford@reading.ac.uk

Koen Bostoen  
KongoKing Research Group  
Department of Languages and Cultures: Africa  
Ghent University  
9000 Ghent, Belgium  
koen.bostoen@UGent.be
Résumé

Durant le XXᵉ siècle, de nombreuses études lexicostatistiques ont été réalisées sur les langues bantu permettant ainsi une meilleure compréhension de celles-ci. Par la suite, le développement des méthodes de classification a permis d’appliquer aux données linguistiques les méthodes appelées « phylogénétiques » empruntées au domaine de la biologie. Pour cette nouvelle étude, nous avons sélectionné le lexique de base de 95 langues bantu représentatives des zones nord-ouest et ouest, avec une attention particulière portée sur la région du Bas-Congo. Les résultats mettent en exergue la découverte d’un groupe particulier que nous avons nommé « agglomérat linguistique kikongo » (Kikongo Language Cluster – KLC), qui se situe au sein du groupe bantu de la côte ouest. Ce groupe, composé de langues bantu proches, constitue cependant un groupe hétérogène. Afin de mieux définir la vraie nature du KLC, nous avons appliqué à nos données les méthodes phylogénétiques. Les arbres et réseaux obtenus permettent de commencer à reconstruire l’histoire et l’évolution de cet agglomérat linguistique.