10 How varied typologically are the languages of Africa?

Michael Cysouw and Bernard Comrie

10.1 Investigating typological variety

Our aim in this chapter is to investigate to what extent it is possible to pick up signals of prehistoric events by studying the distribution of typological diversity across the languages of Africa. The chapter is experimental, in the sense that it aims to test a particular method rather than to assume that the method in question is valid. The results of the investigation will show that, while there are clear limitations especially as one goes further back into history, nonetheless there are clear signals of prehistoric events that can be traced in the geography of typological diversity in Africa. Our aim is not to develop a method that will replace other methods, in particular the comparative method in historical linguistics (Campbell 2004), but rather to see what contributions can be made in specific areas by other methods, in this particular case areal typology.

When we speak of the geographical distribution of typological diversity, we are concerned with typological or structural features of languages, for instance whether they have phonemic tone or not, whether in their basic constituent order the attributive adjective precedes the noun or follows it, etc. Crucially, we are concerned with the extent to which languages are typologically, i.e. structurally, similar to one another or different from one another. Until recently, judgments of the typological distance between languages have been largely subjective, or restricted to a very small set of typological parameters. This situation has changed substantially with the publication of the *World Atlas of Language Structures* (Haspelmath et al. 2005, hereafter WALS). WALS provides detailed information on the geographical distribution of over 130 structural features across the languages of the world. The project relies on a basic sample of 200 languages, although for some features the relevant data for a particular language are missing, while for others data are provided for more than the 200 languages of the basic sample. WALS comprises both a printed atlas and an online version WALS. info, the latter being particularly useful for carrying out linguistic research.

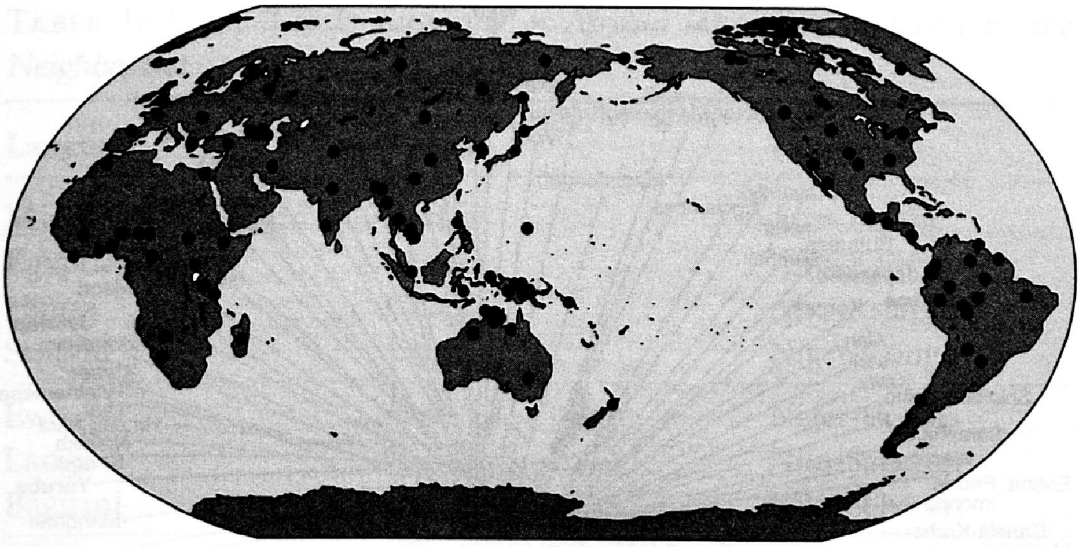
Using WALS, it is possible to measure the typological distance between two languages, essentially by calculating the number of structural features on which the two languages differ in value relative to the total number of structural features for which WALS provides data on both languages (known as the "relative Hamming distance" in biology or the "relativer Identitätswert" (RIW) in dialectology, Goebl 1984). Thus, if the number of features treated remains constant, a pair of languages will be typologically closer the more feature values they have in common, and typologically more distant the fewer they have in common. For the purposes of this exploratory study, we have not made any attempt to weight features differently. Although this is technically easily possible, it is not obvious on which basis linguistic features should be weighted (for attempts to establish weights of WALS features, see Wichmann and Kamholz 2008 for weights related to diachronical stability; and Cysouw et al. 2008 for weights related to the overall typological profile). Further, as noted above, there is the problem that WALS has a rather unevenly filled data table. Different languages that occur in WALS may occur in the treatment of more or fewer structural features. In order to maintain statistical reliability, we restrict ourselves in our various samples in this chapter to languages for which data on a sufficient number of structural features are available.

10.2 Africa in relation to the rest of the world

The first question that we pose is whether the languages of Africa, taken as a whole, form anything like a typological grouping, i.e. a set that is internally relatively homogeneous but also relatively distinct from languages spoken in other parts of the world.

10.2.1 *Africa and the whole world*

For this purpose, we first constructed a worldwide sample of 102 languages from WALS, as shown in Map 10.1. These 102 language are chosen



Map 10.1 A worldwide sample of 102 languages from WALS.

by selecting, first, the languages with the most available data points from each genus to avoid bias stemming from closely related languages.[1] Second, we restricted this "best per genus" sample rather arbitrarily to the 100 best coded languages, but ended up with 102 because various languages had the same number of available data points. We then constructed a NeighborNet (Bryant and Moulton 2004) expressing the degree of typological distance among the languages in the sample, shown in Figure 10.1. In this figure, similar languages are placed closer to each other, sharing parallel lines to the extent that they share linguistic similarities. However, languages are not forced into groups (as is the case in many other clustering algorithms), giving a visual impression of the amount of evidence for many alternative grouping.

The resulting network shows little internal structure, with nearly all languages being at the end of long lines unique to that language. Only few smaller groups of languages are discernible. This indicates that from a worldwide perspective, the structural characteristics from WALS do not show strong evidence for larger subgrouping of languages. Moreover, no distinctively African grouping emerges (the African languages are

[1] A genus—plural: genera—is a group of languages whose genealogical relatedness is visible by inspection, corresponding to a time depth of up to 2,500–3,000 years, roughly equivalent to the major branches of the Indo-European family, like Germanic or Romance.

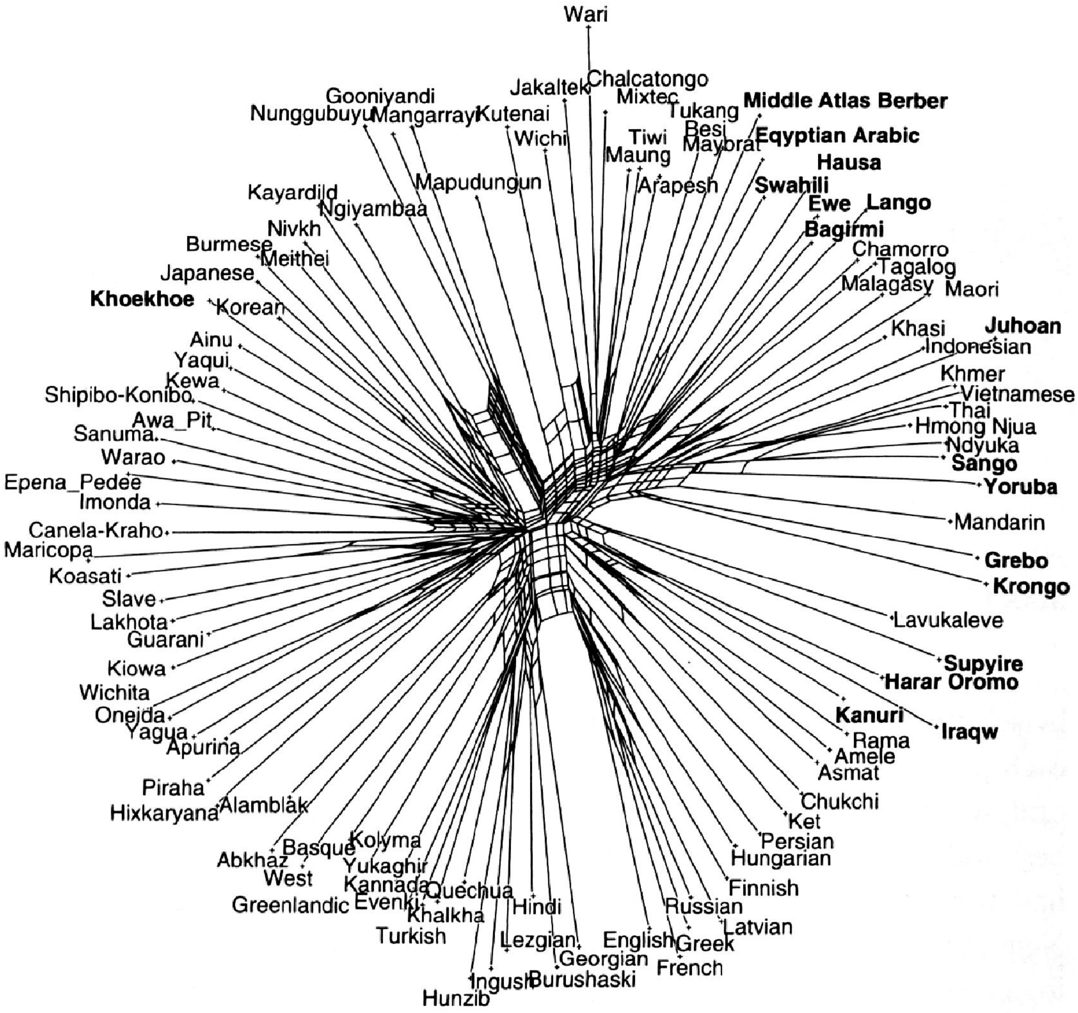


FIG. 10.1 NeighborNet of the 102-language sample; languages from Africa are in bold type.

shown in a larger and bold typeface in the figure). One African language, Khoekhoe, is placed very distant from the others, and even among the part of the network that contains the other African languages there are many intervening non-African languages, i.e. it is not uncommon for an African language to be closer to some non-African language than to some African language. Looking more closely at the smaller-scale clustering of African languages, various clusters are discernible, as summarized in Table 10.1. All these languages are from different genera, because this was one of the grounds on which the languages were chosen. However, even from a deeper genealogical perspective these groups do not show any consistent historical profile (shown in Table 10.1 are the large-scale African families Afro-Asiatic, Niger-Congo, and Nilo-Saharan as proposed by Greenberg 1963).

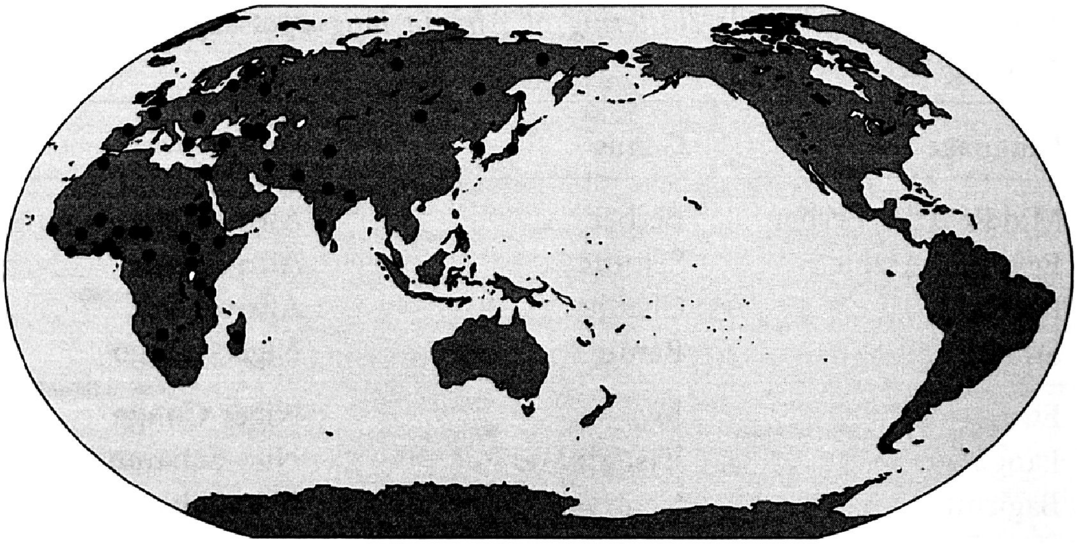
TABLE 10.1 *Small-scale clustering of African languages suggested by the NeighborNet.*

Language	Genus	Family
Middle Atlas Berber	Berber	Afro-Asiatic
Egyptian Arabic	Semitic	Afro-Asiatic
Hausa	Chadic	Afro-Asiatic
Swahili	Bantu	Niger-Congo
Ewe	Kwa	Niger-Congo
Lango	Eastern Sudanic	Nilo-Saharan
Bagirmi	Central Sudanic	Nilo-Saharan
Sango	Adamawa-Ubangian	Niger-Congo
Yoruba	Defoid	Niger-Congo
Grebo	Kru	Niger-Congo
Krongo	Kadugli	(disputed/unknown)
Kanuri	Saharan	Nilo-Saharan
Iraqw	Southern Cushitic	Afro-Asiatic
Harar Oromo	Eastern Cushitic	Afro-Asiatic

10.2.2 *Africa and Eurasia*

Second, we carried out essentially the same procedure again, but this time restricting ourselves to languages of Africa and Eurasia, with the 56-language sample illustrated in Map 10.2. The reason for this restriction is that we expect to find more structure in the language similarity when looking at continent-sized areas. Indeed, the resulting NeighborNet in Figure 10.2 shows considerably more structure than did Figure 10.1 and it reveals rather clearly an African clustering (the African languages are shown in a larger and bold typeface in the figure), though Khoekhoe is still in an isolated position relative to the other African languages.

However, before interpreting this result too far, we need to consider other factors. In particular, we know from other studies based on WALS (cf. Cysouw 2006) that typological distance correlates highly with geographical distance, i.e. that languages spoken in the same neighborhood tend to be typologically more similar to one another than languages spoken further apart. For the 56-language sample, this correlation is



Map 10.2 A sample of 56 languages, restricted to Africa and Eurasia.

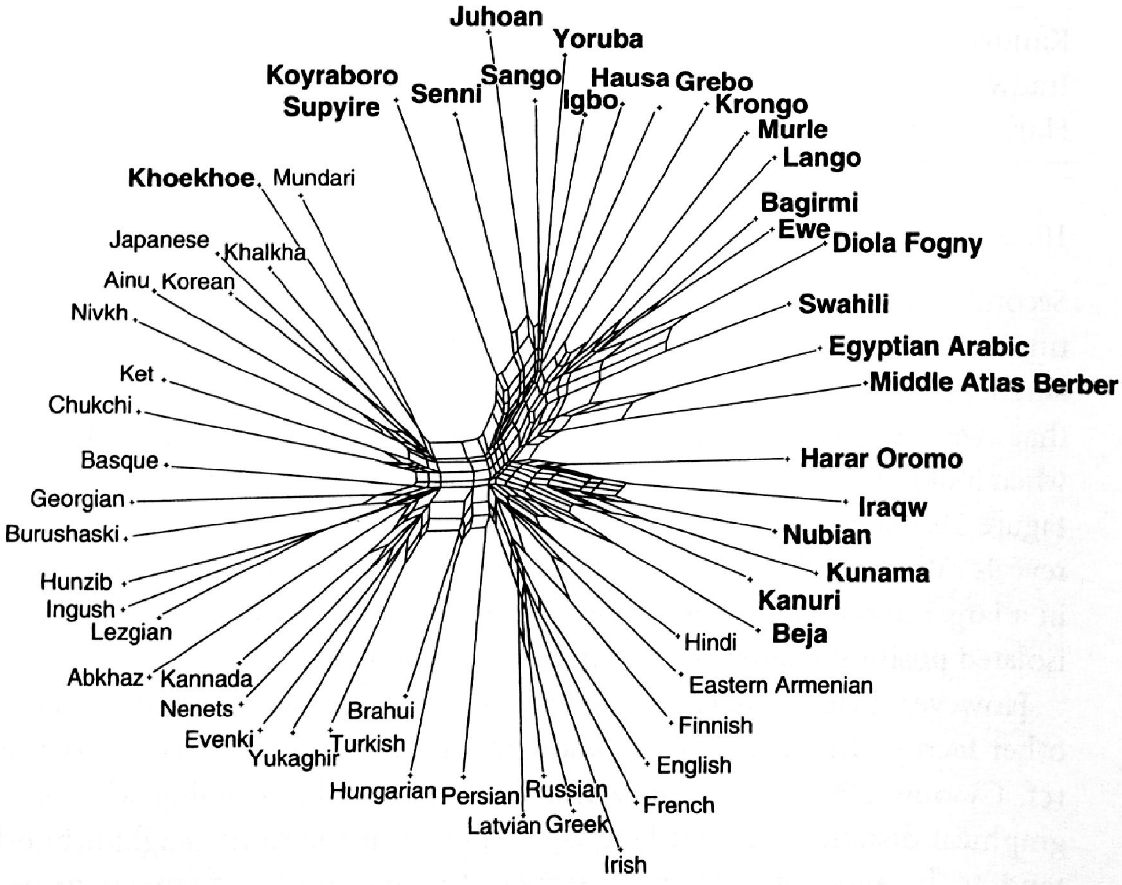


Fig. 10.2 NeighborNet of 56 languages, restricted to Africa and Eurasia.

shown in Figure 10.3, which plots typographical distance against geographical distance. There is a reasonably strong, and clearly significant, correlation between geographical distance and typological distance (Pearson's $r=.39$, Mantel Test $p<.0001$). As Africa and Eurasia are geographically nicely separated, at least part of the distinction between African and Eurasian languages as found in Figure 10.2 can be explained by geographical distance.

There are various possible explanations for such a significant correlation between geography and linguistic structure. We favor an interpretation that gives prominence to horizontal transfer (i.e. borrowing). In contrast to biological diversification in the animal kingdom, horizontal transfer plays a very significant role in the history of language. We would like to suggest that the attested correlation between geography and typology is caused to a large extent by convergent evolution through borrowing (which is more likely to happen between geographically close languages). In the case of our language sample an alternative "isolation by distance" approach does not seem fruitful. First, the sample consists of languages that are not obviously related, so any spreads must have been very long ago, and, second, we are talking about massive geographical distances. Finally, note that relatively recent spreads of languages would

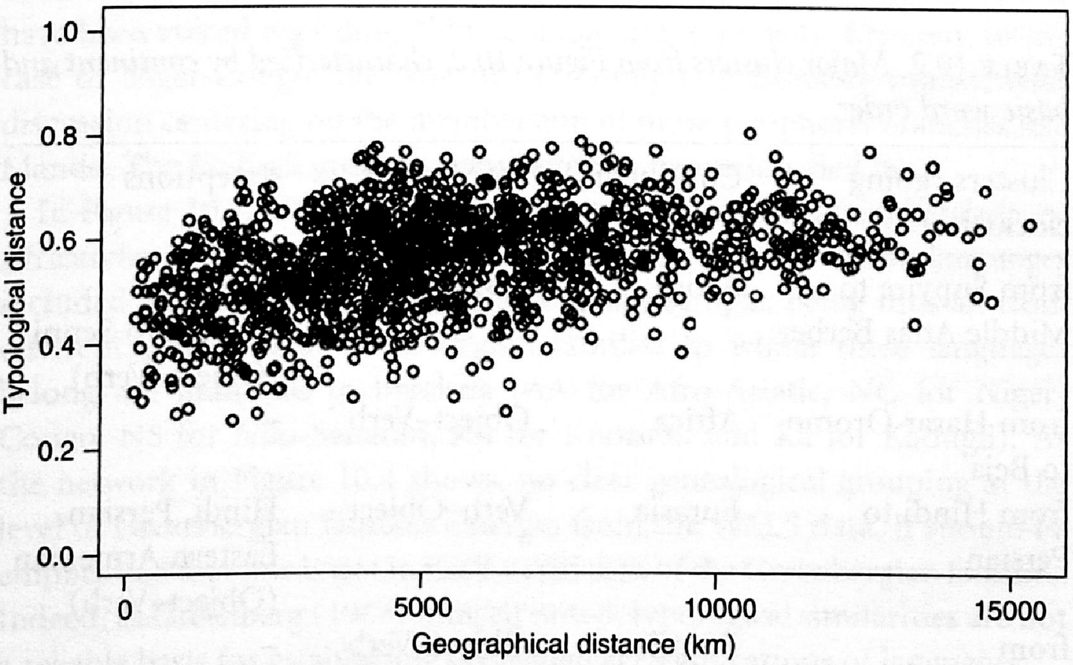


FIG. 10.3 Correlation between geographical distance and typological distance for all pairs of languages from the 56-language sample.

actually result in a less pronounced trend, as even far-away languages would still show strong similarities due to their common origin (cf. the case of the Bantu expansion discussed in section 10.3.2).

Looking somewhat more closely into the groupings discernible in Figure 10.2, it looks like the African languages (apart from Khoekhoe) are separated into two groups (going clockwise through the network: the first cluster ranging from Supyire to Middle Atlas Berber and the second ranging from Harar Oromo to Beja). Likewise, the Eurasian languages also seem to have a major division into two groups (the first cluster ranging from Hindi to Persian, the second from Turkish to Mundari; Brahui and Hungarian being somewhere in the middle). These two separations almost perfectly correlate with the order of object and verb (Dryer 2005) as summarized in Table 10.2. We found the same strong impact of the order of object and verb on the overall typological similarities also in another study based on the WALS data, in that paper focusing on the languages from New Guinea (Comrie and Cysouw forthcoming).

As an interim summary, we may say that in terms of a whole-world comparison, the languages from Africa do not emerge as a typologically distinct subgroup. With respect to the comparison of Africa and Eurasia, things seem to be better, with a clear African subgroup. However, this is at least partially caused by geographical proximity.

TABLE 10.2 *Major clusters from Figure 10.2, characterized by continent and basic word order.*

Clusters (going clockwise)	Continent	Word order	Exceptions
from Supyire to Middle Atlas Berber	Africa	Verb–Object	Supyire, Koyraboro Senni (Object–Verb)
from Harar Oromo to Beja	Africa	Object–Verb	–
from Hindi to Persian	Eurasia	Verb–Object	Hindi, Persian, Eastern Armenian (Object–Verb)
from Turkish to Mundari	Eurasia (+ Khoehoe)	Object–Verb	–

10.3 Relations among African languages

We now turn more specifically to internal relations among the languages of Africa. We proceed as follows. First we consider large-scale genealogical groupings of languages, called language families, which represent a considerable time depth. We then turn to lower-level genealogical groupings, namely genera, which reflect a shallower time depth. In each case, we pose the following question: Are members of pairs of languages within the given genealogical grouping more similar to one another than members of pairs of languages across the relevant genealogical boundary? The answer to this question is then tested against geography, to check whether the patterning could be the result of geographical proximity rather than typological similarity.

10.3.1 *Language families*

To investigate typological diversity within and across language families, we work basically with the four language families posited by Greenberg (1963), namely Afro-Asiatic, Nilo-Saharan, Niger-Congo (the more usual current term for Greenberg's Niger-Kordofanian), and Khoisan. We are, of course, aware that not all of Greenberg's classification is considered robust within African linguistics. In particular, serious doubts have been voiced regarding Nilo-Saharan and, especially, Khoisan. In the case of Niger-Congo, the core of the family is reasonably robust, with discussion centering on the membership of more peripheral branches like Mande. The Kadugli group is considered to be unclassified here.

In Figure 10.4, a NeighborNet of the typological distances between 24 African languages is shown. These 24 languages are the African languages included in the previously used 56-language sample, being thus all from different genera. The Greenbergian families to which these languages belong are indicated in brackets (AA for Afro-Asiatic, NC for Niger-Congo, NS for Nilo-Saharan, Kh for Khoisan, and Ka for Kadugli). As the network in Figure 10.4 shows, no clear genealogical grouping at the level of Greenbergian families emerges from the WALS data. It should be emphasized that this is not in itself a criticism of the Greenbergian families. Indeed, as Greenberg (1963) himself noted, typological similarities are not a reliable basis for establishing genealogical classifications of languages.

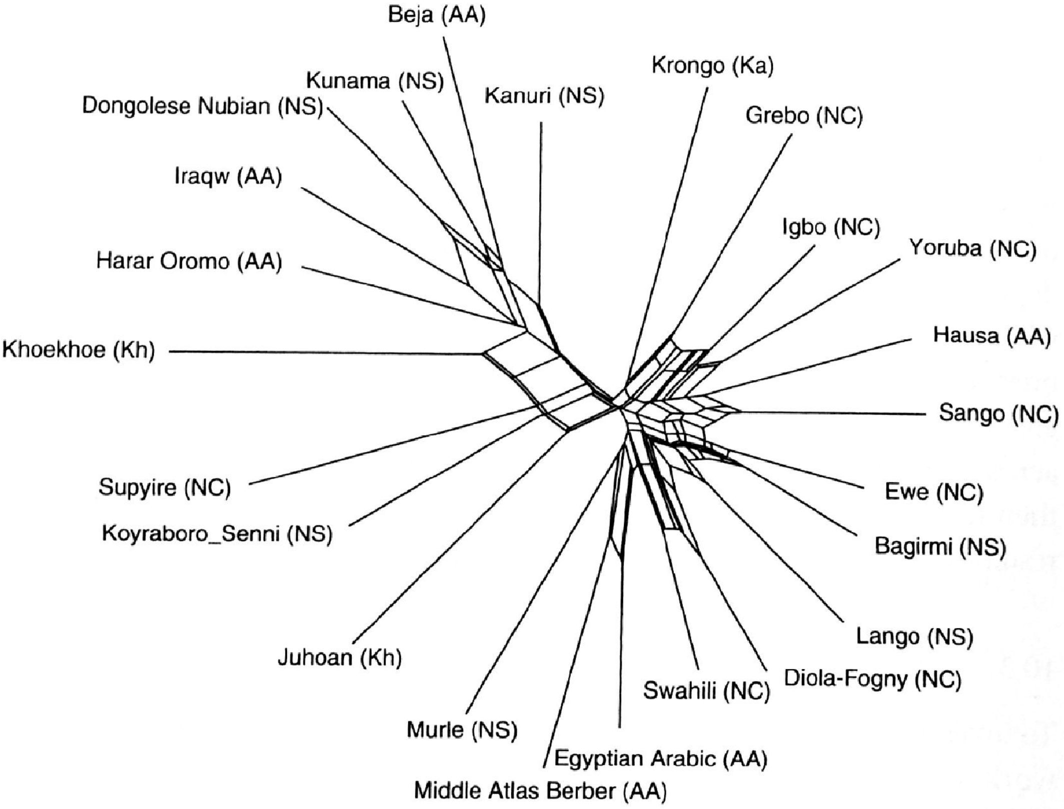


FIG. 10.4 NeighborNet of the 24 languages from Africa in the sample.

Actually, there is a (small) signal of the Greenbergian families to be found in the WALS data. Figure 10.5 shows the significant correlation between geographical distance and typological distance for the 24 African languages (Pearson's $r=.33$, Mantel Test $p<.0001$). The two lines in the figure are the regression lines for the pairs of languages that are not related (upper line) and the pairs of languages from the same Greenbergian family (lower line). As can be seen, there is a slight tendency for the languages from the same family to be typologically more similar (viz. the line is lower) and to be less dependent on geographical distance (viz. the line is less steep). However, the differences are very small (regression of upper line: $typ=.48+1.9{\cdot}10^{-5}$ (*geo*) vs. lower line $typ=.46+7.8{\cdot}10^{-6}$ (*geo*).

10.3.2 *Language genera*

For the investigation of the impact of lower-level genealogical groupings on typological distances among African languages we constructed a sample including multiple languages from the same genus, while still keeping

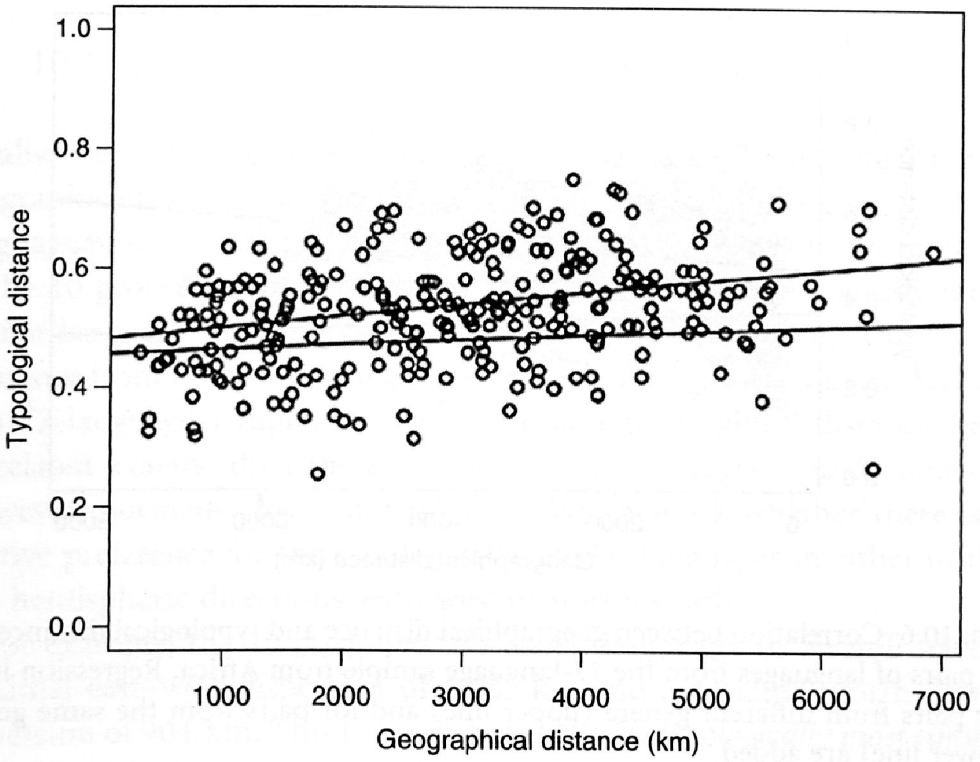


FIG. 10.5 Correlation between geographical distance and typological distance for all pairs of languages from the 24 African languages. Regression lines for pairs from different families (upper line) and for pairs from the same families (lower line) are added.

to a decent amount of available data for all the languages in the sample. Given the data as provided by WALS, it turns out to be impossible to meet both these desiderata. In the end we decided to include more languages (partly with less available data) to be able to investigate within-genera against between-genera diversity. The chosen sample of 77 African languages includes languages from 37 different genera, of which 17 genera are represented by more than one language. To reach such a coverage, the lower boundary for data availability per language had to be set as low as 30% of the available features in WALS.

Figure 10.6 again plots typological distance against geographical distance for all pairs of languages from the 77-language sample. The regression line for languages from different genera (upper line) is now clearly distinct from the regression line for languages from the same genus (lower line). The precise values are: regression of the upper line: $typ = .47 + 1.8{\cdot}10^{-5}$ (*geo*),

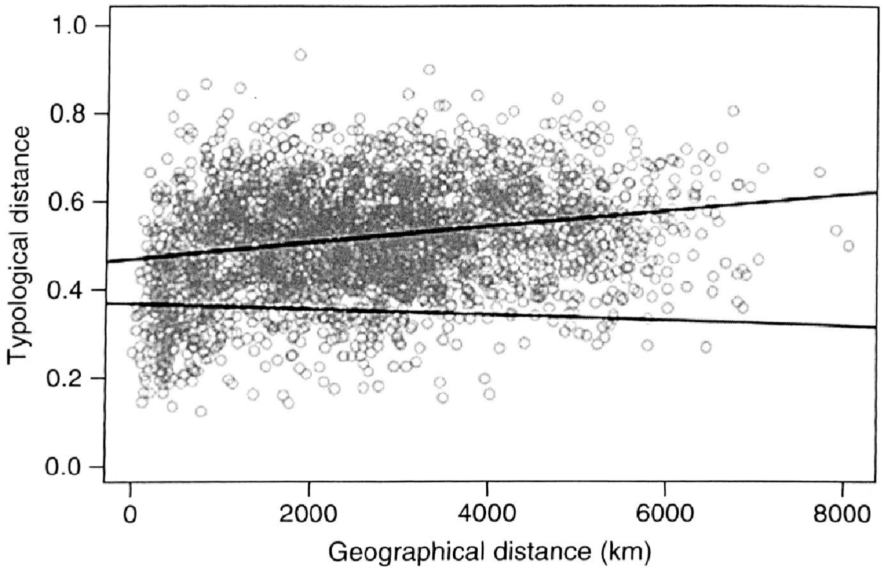


FIG. 10.6 Correlation between geographical distance and typological distance for all pairs of languages from the 77-language sample from Africa. Regression lines for pairs from different genera (upper line) and for pairs from the same genus (lower line) are added.

and regression of the lower line: $typ = .37 - 6.6 \cdot 10^{-6}$ (geo). Typological similarity for a given geographical distance is more likely for languages within the same genera than for languages from different genera, i.e. here we do have a clear signal in the geographical distribution of typological diversity corresponding to the time depth at which genera were formed.

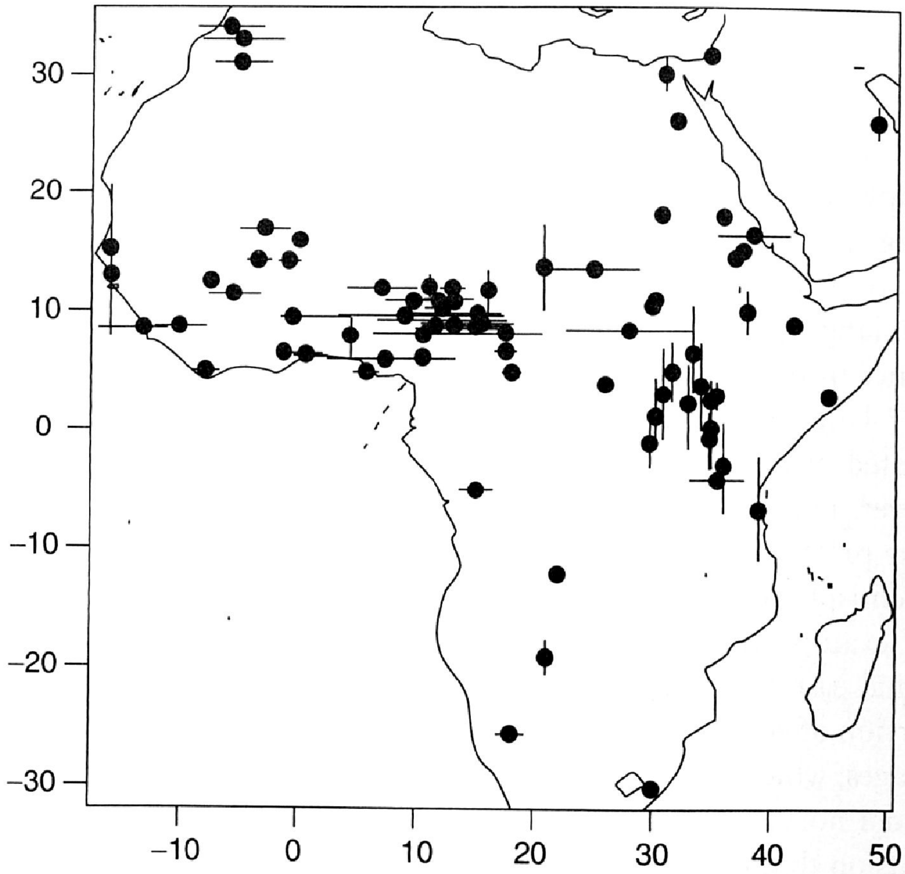
It is important not to overlook one surprising feature, represented by the cline of the lower regression line, namely that for languages within the same genus, typological similarity actually seems to increase with greater geographical distance. On closer inspection, this appears to be caused by the relatively recent expansion of the Bantu languages. The few genera that are more widespread geographically, such as Semitic and Bantu, are the result of recent expansions, which accounts for the large geographical distances relative to rather low typological distance. If one takes two African languages that belong to the same genus but are spoken very far apart, they are almost certain to be Bantu languages, and Bantu languages are typologically very similar to one another. In other words, the clearest signal we may be receiving here is of the Bantu expansion. When removing the Bantu genus from the regression, it becomes $typ = .37 + 9.3 \cdot 10^{-6}$ (geo), now with a positive cline.

10.4 Geographical distribution of typological variety

Finally, we turn to the question: How is typological variety distributed geographically in Africa? We answer this question by considering, for each language in the 77-language sample, the relation between the distribution of the 10 geographically closest languages and the 10 typologically most similar languages (the number 10 is chosen to be large enough to prevent influence from random factors, but small enough to lead to differentiation in a 77-language sample). If typological and geographical distance were correlated exactly, then these two sets of 10 languages would coincide. However, normally this is not the case. We then ask whether there is a relative preference for the typologically similar languages in either of the two hemispheric directions, east–west or north–south.

For example, the 10 *geographically closest* languages to Swahili have a maximal east–west dispersion of 1,685 km and a maximal north–south dispersion of 904 km. This is compared to the 10 *typologically most similar* languages, which have a maximal east–west dispersion of 2,222 km and a maximal north–south dispersion of 1,976 km. These figures lead to the conclusion that the typologically most similar languages have a relatively larger spread in the north–south direction, because 1,976/904 = 2.2 is larger than 2,222/1,685 = 1.3. In other words, Swahili shows a north–south preference for typological similarity, presumably reflecting greater language contact to the north and south, leading to more spread of linguistic features along that axis.

Map 10.3 extends this to all languages in the sample, showing that some languages have more relative typological similarity in the east–west direction (represented by horizontal lines), others more relative typological similarity in the north–south direction (represented by vertical lines). The length of the lines is representative for the strength of the preference in either direction (ultimately leading to a point when there is no preference for either direction). Interestingly, there are parts of Africa, such as west-central Africa, that are predominantly "horizontal," and other parts, like east Africa, that are predominantly "vertical." We propose the following tentative explanation for this distribution (following Güldemann 2008 and forthcoming b). The horizontal preference may reflect the greater ease of population movements and contacts on an east–west axis given that climatic and related changes are most significant on the



MAP 10.3 Hemispheric preference for typological similarity for the 77-language sample of African languages.

north–south axis, i.e. migration and contact to the east or west minimizes the need to adapt to new climates, as argued, for example, by Diamond (1998). But under particular circumstances, movement and contact in a north–south direction may be favored, for instance along an east or west coast (cf. the west coast of Africa in Map 10.3), or in an area where valleys running north–south are separated by mountain ranges impeding east–west movement, as with the Rift Valley in East Africa.

10.5 Conclusions

Our investigations show that, based on the typological data from WALS, there is no genealogical signal in the African languages as a whole, nor in the typological variety of the Greenbergian language families. Further, typological similarity is correlated with geographical proximity, which indicates that typological similarity is (possibly strongly) influenced by borrowing (i.e. horizontal transfer). At the level of genera, things are somewhat better for signals from vertical transfer, though still not perfect. In particular, signals of particular recent events, like the Bantu expansion, seem to play a major role in the distribution of typological similarity. Further, we have argued that linguistic similarity, whatever its cause, does show clear geographical structure within Africa.

Being able to identify how typology, geography, and history interact, including being able to identify our limitations in evaluating this interaction, provides an important supplementary insight into prehistoric demographic processes. Our investigation also serves as a warning: Application of statistical methods to data based on typological properties of language may well reflect relatively recent historical events; but it seems rather that such features cannot be used to reconstruct more distant events, such as at the level of large-scale language families. We hope, nonetheless, that further refinement of these methods will lead to finer discrimination of different chronological layers.