Geographical axis effects in large-scale linguistic distributions

Tom Güldemann (HU Berlin and MPI-EVA Leipzig) and Harald Hammarström (RU Nijmegen and MPI for Psycholinguistics Nijmegen)

Abstract

Taking up Diamond's (1999, chapter 7) idea regarding the different population histories of continental areas, Güldemann (2008, 2010) speculatively proposed that macro-areal aggregations of linguistic features might be influenced by large-scale geographical factors. In line with Diamond's geographical axis hypothesis, it is assumed that the way linguistic features assemble over long time spans and large geographical space is determined among other things by two factors which can be called the "latitude spread potential" and the "longitude spread constraint". This paper reports on first results of testing this idea. With respect to the first factor we argue in particular that contact-induced feature distributions as well as genealogically defined language groups that have a sufficient geographical extension tend to have a latitudinal orientation. Regarding the second factor, we provide first results suggesting that linguistic diversity within language families tends to be higher along longitude axes. If these findings can be replicated by more extensive and diverse testing, they promise to become important ingredients for a comprehensive theory of human history across space and time within linguistics and beyond.

1 Introduction: areal linguistics and linguistic areas

Building on the observation that the histories of human populations have been quite different across distinct continental areas, Diamond (1999, chapter 7) argued that in the long term the historical dynamics of continents are decisively determined by the orientation of their geographical axis: spreads of people and their features are facilitated along latitude axes but hampered along longitude axes. This would be due to the fact that climatic-ecological factors remain more homogeneous in a west-east direction and therefore human adaptation and subsistence conditions are usually more similar along latitudes. For example, the same animals and plants can be utilized if people migrate East/West rather than North/South. These two interrelated phenomena, so-to-speak "the two sides of one coin", will be called henceforth the "latitude spread potential" and the "longitude spread constraint".

Trying to interpret his findings on large-scale distributions of linguistic features within the African continent, Güldemann (2008, 2010) proposed to extend Diamond's geographical-axis hypothesis in two directions. First, it should not only be relevant for the spread of cultural artifacts and ideas but also of linguistic features associated with languages

whose speakers migrate and/or are in contact both on the level of individuals and entire communities. Second, if the relevant geographical area associated with a certain feature distribution is sufficiently large, the mechanism should not only leave traces in areas of continental but also smaller sub-continental size.

In other words, the geographical-axis effect (as a statistical tendency concerning numerous individual spreads of populations and their features in space and time) can be expected to be a major factor influencing the formation of large-scale aggregations of linguistic features. This hypothesis needs to be seen before the background of the current general discussion in areal linguistics and in particular within this sub-discipline on the assumed role of geography.

With respect to the first issue, it is important to recognize that in the recent past there has been an intensified debate about the concept of "linguistic area". This can be discerned from the large number of more theoretically oriented papers like, e.g., Muysken (2000, 2008), Thomason (2001: 99), Dahl (2001), Stolz (2002, 2006), Bisang (2006a, b), Campbell (2006), Tosco (2008), Bickel and Nichols (2012), and Muysken et al. (forthcoming). In particular, in a number of these works one can notice a considerable tendency towards a pessimistic look at the very usefulness of this concept. At the same time, it has been realized that a potential way out of the problem is a better integration of the contact-induced linguistic-area concept in a more general theory of areal linguistics. In this framework, the first step would have to focus on the mere fact that/whether a geographical distribution of linguistic features can be observed and is at all somehow significant. Within areal linguistics a linguistic area is viewed then more neutrally as a distribution of features according to a non-trivial "compact" geographical entity independent of any historical (or other) explanation. At this stage, one would be concerned first of all with the statistical probability that an identified distribution is diagnostic of an interesting relation between the feature and the associated area as opposed to mere coincidence. While this issue must be considered essential, it cannot be pursued further in the context of this paper (but see Daumé 2009; Lucas, Cule and Mathieson 2009; Chang, Michael and Stark forthcoming; and Muysken et al. forthcoming).

A linguistic area in the traditional, narrow sense entails a second step: it is a feature distribution according to a compact geographical pattern COUPLED with a specific historical scenario, namely that language contact is the central explanation for the observed distribution, rather than coincidence, universal tendencies, and in particular genealogical inheritance. Henceforth, we will call this particular concept a "(linguistic) contact(-induced) area" in order to avoid confusion between the two senses of linguistic area entertained here.

The second issue of the relation between linguistic areas and geographical space is equally controversial. On the one hand, there is no doubt that at least some large-scale linguistic distributions are partly determined by geographical factors, e.g., the significantly higher linguistic diversity in the tropics (cf. Nettle 1999; Collard and Foley 2002).

Moreover, it is intuitively clear that geographical patterns and events are among the factors which determine linguistic history in space and time, including the presence, trajectory, and speed of contact-induced diffusion, and thus steer more generally the distributional dynamics of features. There are, of course, well attested cases to this effect (cf. such a recent exemplary study as Bostoen, Grollemund and Muluwa 2013). Given this, there is no a-priori reason why the result of some such events should not linger on for a longer time period in the form of a particular spatial distribution of one or more linguistic features.

At the same time, a causal role of geography has recently been denied explicitly for contact-induced areas by Campbell (2006).

> … it is the diffusion that is of prime importance, and … the geographical aspect of putative 'linguistic areas' is derivative.
>
> ... the linguistic borrowings are prime, and the geographical areas are only a reflection of these, with no significant causal force of their own. (ibid.: 16)

Admittedly, this claim may have been intended primarily for intermediate and small contact areas, although even here this strong statement remains to be tested. In any case, the strong claim that geography has a purely "derivative" role in areal linguistics would be an offhand dismissal and the general question deserves to be investigated more systematically.

Intuitively, geography should become more important the larger an areal distribution, whatever its ultimate origin. Inversely, such big areal patterns involve greater time depths and it is thus less likely that one is able to reconstruct the specific historical scenarios or even concrete events and circumstances that caused the synchronic picture. The last point is nicely captured by Muysken's (2008) holistic approach to language contact in general, which includes large contact areas. As seen in Table 1, he introduces several levels of scale which differ in certain parameters, among them the kind of historical scenarios entertained (last column). In this context, we would venture to add an additional dimension, namely the role of geography, which we assume to be small on the micro level and important on the macro level.

Level	**Space**	**Time**	**Sources**	**Scenarios**
Micro	Bilingual community	20-200 years	Fieldwork data	Specific contact scenarios
Meso	Geographical region	Generally 200-1000 years	Comparative data; historical sources	Global contact scenarios
Macro	Larger areas of the world	Deep time	Typological, genetic, archeological data	Vague or no contact scenarios

Table 1: Levels of scale in the analysis of linguistic contact areas (Muysken 2008: 5)

The way a linguistic feature ends up in a particular language variety that is spoken in a certain geographical location - the individual data point of a large-scale distribution

pattern - is an extremely complex matter, normally without any direct relation between geographical space and feature.[1] The relation is instead mediated by intermediate layers, which requires one to take such diverse phenomena as the following into account:

a) the feature within the linguistic system of an idiolectal variety,

b) the idiolect as a member lect of an abstract language,

c) the language spoken by an abstract population found in a certain location,

d) the location in geographical space.

Movement of features through geographical space largely happens in two idealized ways: a) across geographically "stable" human populations through contact between them (and thus metaphorically between languages) or b) with/on geographically mobile human populations.

Such complex and indirect feature-geography interrelation, and thus the history of large distribution patterns, can only be captured by more abstract, metaphorical modeling. The idea to conceptualize population features similar to potentially contagious viruses on a host has been entertained both outside and within linguistics - compare, e.g., Cullen (2000) and Enfield (2003, 2008), respectively. In the present context, a yet more abstract concept of (linguistic) features as "particles in a liquid/pulp" is possibly even more appropriate.

What we propose here is that the distributional dynamics of linguistic features - the "particles" of the last metaphor - are not only steered by their more "active" inherent properties and/or the properties of their hosts (lects, languages, speakers, populations) but also by the more "passive" reactive interplay of the different kinds of feature hosts with the geographical environment - the "liquid" - in which all these hosts emerge, thrive, and degrade. As introduced above, the relation between feature aggregations and one such environmental factor, geographical axis, will be dealt with in the remainder of this paper.

2 Geographical axis effects

2.1 The macro-areal profile of Africa

In the recent past the work on large-scale linguistic distributions has been intensified considerably. This holds on a global scale (cf., e.g., Haspelmath et al. 2005) as well as for areas of (sub-)continental size (cf., e.g., Comrie (2007) and Enfield (2005, 2011) on Mainland Southeast Asia). For Africa in particular, macro-areal linguistic studies have been resumed in the late 1990s after a considerable break since the early work by Greenberg (1959, 1983) and Heine (1976).

---

[1] An occasional direct relation has, however, been entertained, e.g., by Fought et al. (2004), Ember and Ember (2007), and Everett (2013).

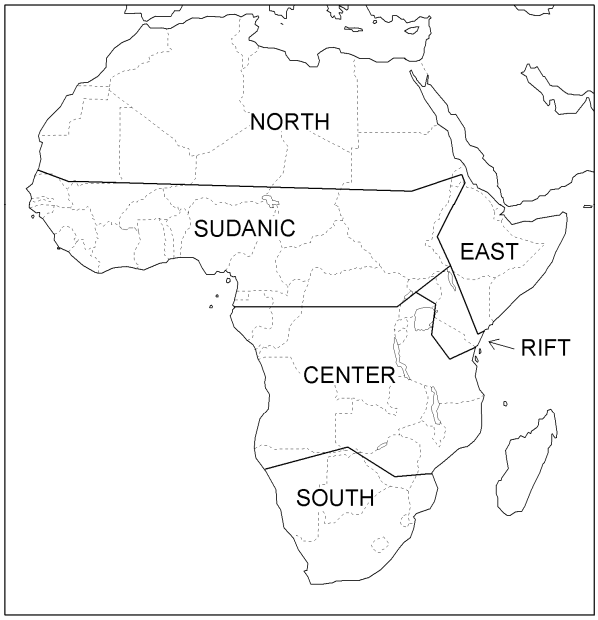
In connection with providing an alternative hypothesis to Greenberg's (1963) poorly substantiated hypothesis of a Khoisan language family, Güldemann (1998, 1999, 2013) proposes that the relevant languages in southern African form a convergence area comprised of three language families, called Tuu, Kx'a, and Khoe-Kwadi (see Güldemann forthcoming). This idea had already been entertained by Greenberg and Heine, but at the time was hard to conceptualize in view of the equally entertained genealogical hypothesis. Two further large feature aggregations induced to a considerable extent by convergence are identified by Güldemann (2003, 2008) and Güldemann (2005) and called Macro-Sudan Belt and Chad-Ethiopia, respectively. These contact areas, too, build on earlier parallel concepts: the Macro-Sudan Belt is prefigured by Greenberg's (1959, 1983) African "core area" and Chad-Ethiopia is geographically essentially the same as one of Heine's (1976) large word order zones for which the name was coined originally.

As mentioned in §1, linguistic macro-areas need not be primarily contact-induced. Africa, e.g., also hosts two linguistically homogeneous zones, because they are predominated by languages that are genealogically related, namely the Sahara spread zone in the north and the Bantu spread zone in the centre.

Taking all the above observation together culminates in a continental profile of five macro-areas presented inter alia in Güldemann (2007, 2010), as shown here in Map 1.

Map 1: Macro-areal profile of Africa according to Güldemann (2007, 2010)

Independently of Güldemann's work, Clements and Rialland (2008) also proposed a set of large, linguistically more homogeneous zones in Africa, based exclusively on phonological criteria. This areal inventory is given in Map 2.



Map 2: Macro-areal profile of Africa according to Clements and Rialland (2008)

The comparison between the Maps 1 and 2, summarized in Table 2, reveals that the independent macro-areal studies[2] resulted in largely similar pictures. Moreover, the characterization of such contact-induced areas as the Macro-Sudan belt and the Kalahari Basin has in the meantime been refined and extended by confirmative data (cf. Idiatov 2010 and Güldemann and Fehn forthcoming, respectively). These two observations suggest so far that the proposed profile for Africa is a robust finding, despite the fact that neither study arrived at their macro-areal division by objective procedures.

Güldemann (2007, 2010)		**Clements and Rialland (2008)**
I Sahara spread zone	=	North
II Chad-Ethiopia	≥	East
not associated	≠	Rift
III Macro-Sudan belt	≤	Sudanic
IV Bantu spread zone	=	Center
V Kalahari Basin	=	South

Table 2: Macro-areal profiles of Africa compared

[2] That the research has indeed been largely independent is evident from the partly non-overlapping data and features, and the fact that none of Güldemann's earlier macro-areal studies other than on the Kalahari Basin are mentioned by Clements and Rialland (2008).

Apart from a somewhat distinct conceptualization of macro-areas, the major differences are threefold: Güldemann's "Macro-Sudan belt" is far more limited in the northeast than Clements and Rialland's "Sudanic" zone; the relevant area is assigned instead to Güldemann's "Chad-Ethiopia" which therefore turns out to be larger than Clements and Rialland' "East" zone; Güldemann identifies a larger area that is not viewed as a macro-area in the positive sense while Clements and Rialland consider it as such calling it "Rift" zone.

These differences do not affect two other interesting observations which hold for both proposals and are relevant for the present topic. First, several areas have a pronounced east-west extension. Second, on a continental scale all areas are distributed according to a horizontal rather than vertical pattern. Taken together, one can conclude that large-scale aggregations of linguistic features currently observable in Africa tend to have an east-west orientation which produces an overall pattern of horizontal layering.

A few remarks on this generalization and the general nature of macro-areas, as conceived here, are in order. First, it should not be expected that macro-areas are eternally stable. Over time they can be subject to change with respect to their size, shape or even very existence, as can be argued for taking the areas of Map 1. For example, Güldemann (2008: 181-2) suggests that the eastern part of the Macro-Sudan belt was affected by the spread of Nilotic languages and thus has shrunk over history. Even more dramatically, areas might dissolve altogether; such a scenario is not too unrealistic for the Kalahari Basin given the effect of Bantu and European colonization leading to the wide-spread extinction and endangerment of most languages which are subsumed under "Khoisan" and represent the core groups of the linguistic contact area.

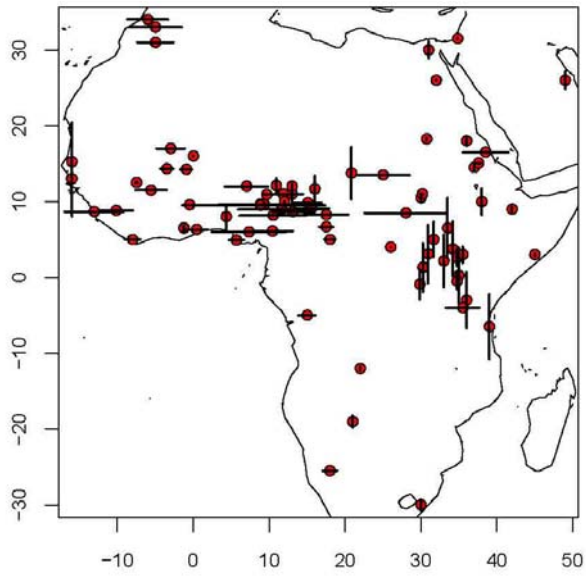
While the most recent macro-areal profile of Africa figures an overall layered pattern of latitudinal areas, it is in principle possible that macro-areas also have a longitudinal orientation. This is due to the fact that the latitude spread potential is not the only geographical let alone general determinant of areal distributions. Notably, other large-scale environmental and geo-morphological features may also shape the movement and sedimentation of linguistic populations and features in space (cf. Nichols 1997: 372-4, 2013). These are notably coastlines, mountain ranges, marked vegetation zones, and water bodies like lakes, rivers, swamps etc. All these can potentially compete with and counteract the impact of the latitude spread potential.

For example, Güldemann (1999, 2010) discusses the possibility that there existed an old longitudinal macro-area in the eastern part of Africa. Synchronically one can identify a dissected distribution of globally quirky obstruent consonant types as well as locally marked head-final traits in the noun phrase which occur in eastern and southern Africa. In between these two zones is situated the eastern part of the large Bantu spread zone in which Bantu languages especially closer to the northeastern and southern periphery show recurrently the rare and Bantu-untypical features of their Non-Bantu neighbors. It can be assumed that such a pattern has emerged from the replacement of Non-Bantu languages that were at least

partly related and similar to those in eastern and southern Africa and that a more compact area across the eastern flank of Africa existed before.

This reconstructed and today submerged area correlates with a longitudinal zone called in geography “High Africa” (cf. Lobeck 1946). “High Africa” is delimited from “Low Africa” in the west by the longitudinal Rift Valley complex comprising several mountainous escarpments in the north and two large lakes in the south; these geophysical landmarks can be argued to have contributed to the formation of a longitudinal linguistic macro-area.

As discussed by Güldemann (2010: 581-2), a first, yet limited study of African language data from the WALS by Cysouw and Comrie (2009) investigating typological diversity across space seems to confirm the assumption that different partly conflicting geographical factors play a role for the geographical distribution of linguistic features. The authors analyzed the geographical orientation of typological similarity~distance in Africa with respect to the WALS language sample represented by 30% or more of the features. For each language they calculate a similarity axis, north-south or east-west, as follows: a given language is pooled with the 10 languages most similar typologically and the 10 languages closest geographically. The geographical dispersion rectangles of these two sets of languages are compared. If the 10 most similar languages are not the same as the 10 geographically closest languages, it can be measured whether typological similarity has an east-west or north-south orientation. The results, shown in Map 3, allow two observations: first, a signal of latitudinal orientation of linguistic similarity concentrated in the Macro-Sudan belt, and second, a signal of longitudinal orientation of linguistic similarity along the Great Rift valley. Although at first glance a complex random picture, these two spatial patterns are in line with what can be expected from geographical circumstances in the relevant areas.



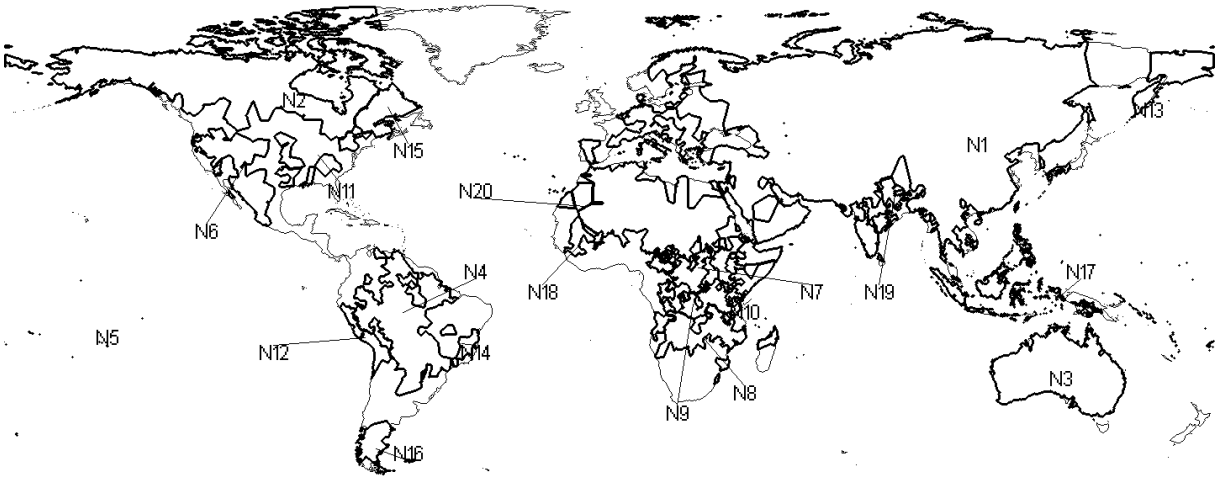
Map 3: Axis orientation of typological similarity in Africa (Cysouw and Comrie 2009)

2.2 Large-scale feature distributions involving linguistic convergence

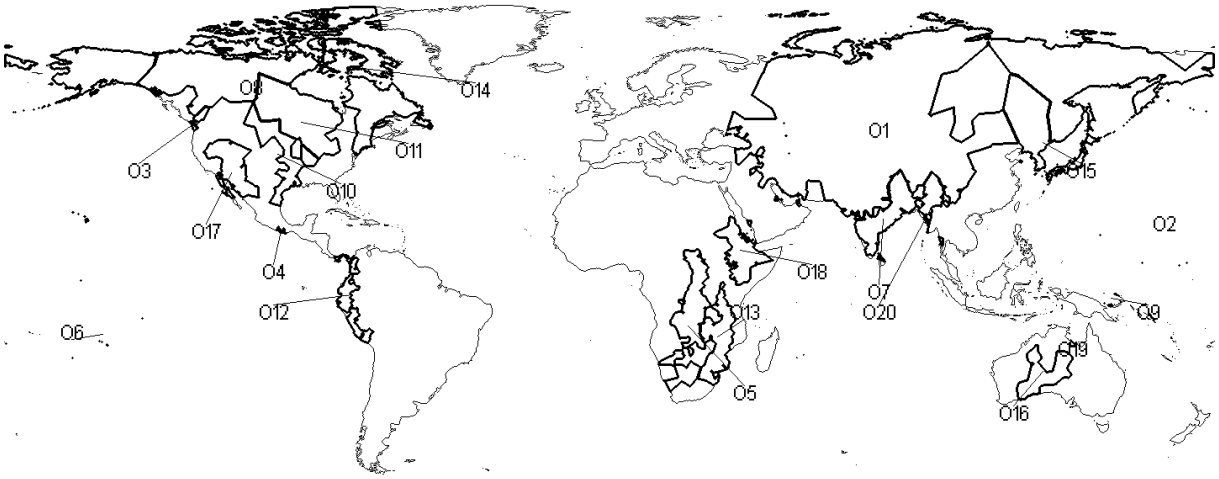
The hypothesis that macro-areas induced by multiple language contact tend to have an east-west axis as a partial reflex of the latitude spread potential is at present hard to test systematically. This is because there is no inventory of such areas that is large-enough for statistical testing and would find a sufficient amount of agreement in the linguistic community. For the time being, a way out of this dilemma is to investigate compact distributions of single linguistic features, irrespective of whether these overlap with other distributions, and determine whether they show any bias with respect to their geographical axis. If geography does play a role, at least some signal should be discernible.

A first attempt in this direction is contained in Hammarström and Güldemann (submitted). The empirical basis is two world-wide data collections: one is on numeral systems containing data for 6837 languages (Hammarström 2010); the other one is on basic word orders in the transitive sentence representing 4653 languages (superseding the data of Dryer 2005, Hammarström 2007, and Lewis et al. 2013).

These two data sets were subjected to various tests one of which yielded relevant results for the geographical-axis hypothesis. The most direct evidence comes from the assessment of areas which are homogeneous with respect to a certain feature value within the two linguistic domains. Apart from gaps in the language coverage, the linguistic data are encoded geographically by means of the center point coordinates of the relevant languages. By joining geographically neighboring languages which have the same feature value and assigning each point on a map to its nearest language we obtain single-feature coherent areas. For details on this procedure the reader is referred to Hammarström and Güldemann (submitted).



Map 4: The 20 largest single-feature areas for numeral systems



Map 5: The 20 largest single-feature areas for transitive sentence order

ID	Numeral systems					ID	Transitive sentence order				
	Type	# lgs	d_{ew}	d_{ns}	d_{ew}/d_{ns}		Type	# lgs	d_{ew}	d_{ns}	d_{ew}/d_{ns}
N1	DECIMAL	2225	19524	10538	1.85	O1	SOV	364	13643	5778	2.36
N3	RESTRICTED	407	4076	3830	1.06	O7	SOV	188	1911	2359	0.81
N4	RESTRICTED	260	3219	4364	0.74	O20	SOV	96	706	1825	0.39
N7	QUINARY	232	3117	2033	1.53	O5	SVO	90	1204	4419	0.27
N9	DECIMAL	147	2872	1945	1.48	O18	SOV	70	923	1869	0.49
N19	DECIMAL	98	881	745	1.18	O12	SOV	68	1114	2709	0.41
N8	QUINARY	89	2580	2283	1.13	O13	SVO	58	924	3100	0.30
N6	QUINARY	76	2798	3113	0.90	O9	SVO	51	1357	2349	0.58
N10	DECIMAL	57	1557	2522	0.62	O4	VSO	33	1458	4965	0.29
N2	QUINARY	54	5946	2781	2.14	O17	SOV	26	1241	1507	0.82
N12	DECIMAL	41	1415	1873	0.76	O2	SVO	24	5141	3325	1.55
N18	VIGESIMAL	24	1160	572	2.03	O10	SOV	24	1191	2578	0.46
N17	QUINARY	24	576	1280	0.45	O8	SOV	13	2373	1375	1.73
N11	DECIMAL	22	2163	1741	1.24	O11	NODOM	13	1610	1892	0.85
N5	DECIMAL	21	5149	2085	2.47	O16	SOV	11	1245	1515	0.82
N14	RESTRICTED	14	979	1165	0.84	O6	VSO	10	3526	1426	2.47
N16	RESTRICTED	7	717	1345	0.53	O3	VSO	8	3038	3461	0.88
N15	QUINARY	7	897	1215	0.74	O15	SOV	7	946	2497	0.38
N13	QUINARY	5	927	1320	0.70	O14	NODOM	7	1140	2278	0.50
N20	VIGESIMAL	3	388	1593	0.24	O19	VSO/VOS	2	4722	347	13.60

Table 3: The 20 largest areal aggregations of the two linguistic features and their geospatial extensions

The Maps 4 and 5 and Table 3 present a subset of the resulting linguistically homogeneous areas, namely the 20 largest ones (in terms of number of languages) for the two feature domains. In accordance with our above hypothesis, we expect that the larger the areas the more their geographical shapes tend to be latitudinal rather than longitudinal. In order to measure the axis orientation we take an area's east-west and north-south endpoints

to get a distance east-west d_{ew} and a distance north-south d_{ns} in kilometers. An area's axis ratio is the ratio of d_{ew}/d_{ns}, whereby a value >1 means that the relevant area is more latitudinal. While there is a lot of variation in the axis ratio across the individual areas, correlating such axis ratios to geospatial size one can determine a mean of axis ratios for all areas that are geospatially larger than a certain threshold size.

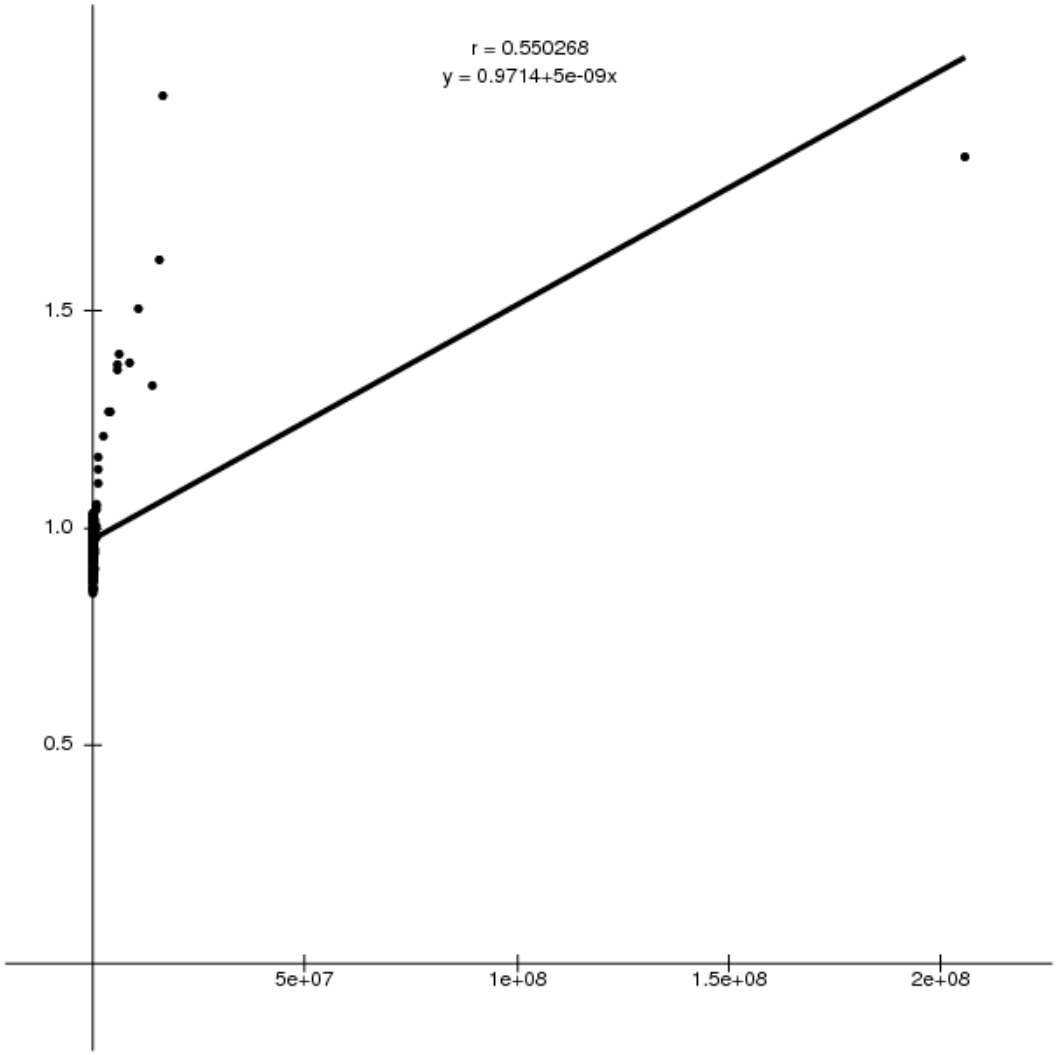


Figure 1: Threshold size-axis ratio plot for numeral bases with fitted regression line

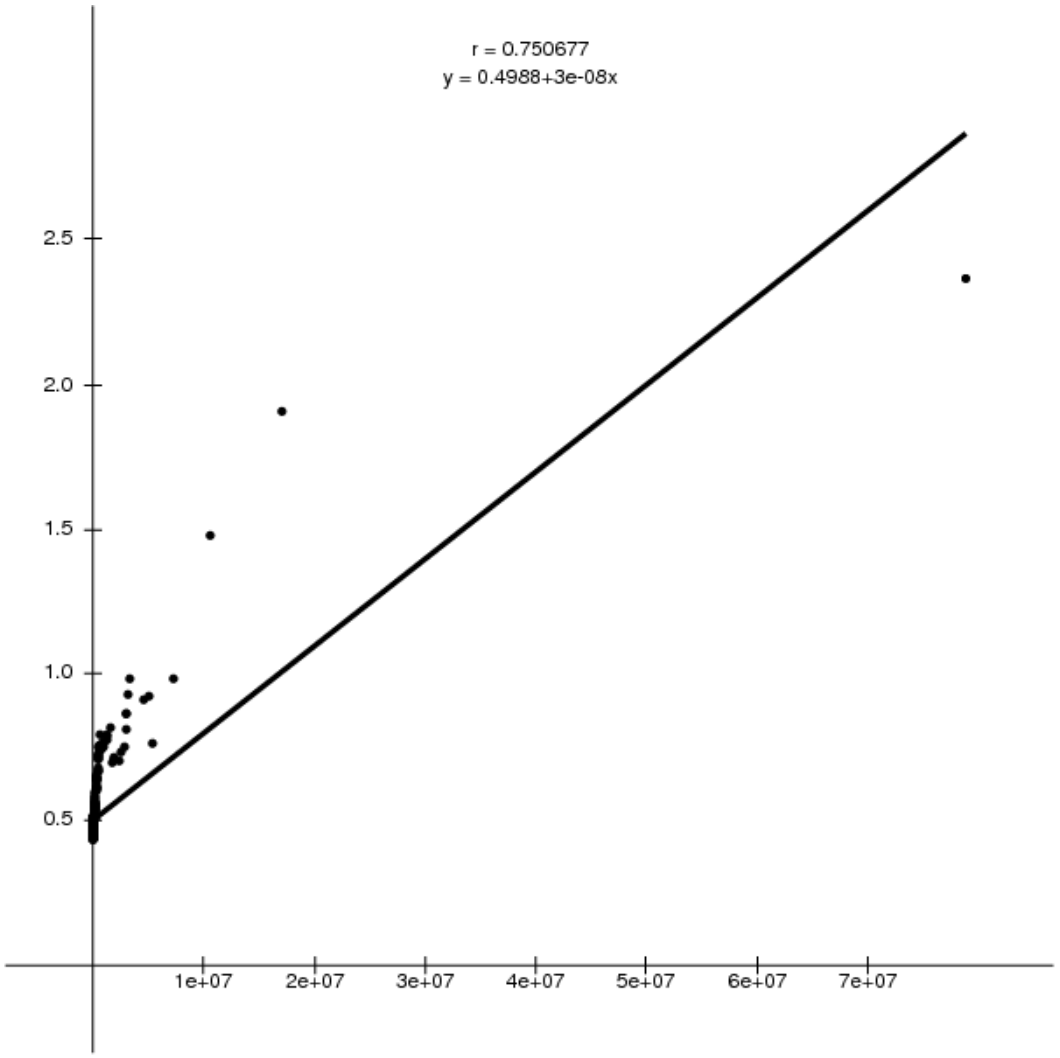


Figure 2: Threshold size-axis ratio plot for transitive sentence word order with fitted regression line

Figures 1 and 2, for numeral bases and transitive sentence word order, respectively, show diagrams of the geometric mean axis ratio (y-axis) plotted against threshold size (x-axis). In both cases, there is a trend such that the mean axis ratio increases with growing threshold size. This can be taken as first positive evidence for the latitude spread potential to influence contact-mediated feature aggregations in geographical space.

2.3 Language families in large geographical space

2.3.1 The latitude spread potential and language family axes

If the latitude spread potential is indeed a factor for large-scale linguistic distributions, it can be expected to have yet further implications. The logic of the hypothesis predicts that the east-west trend holds for any type of historically mediated linguistic distribution which has a sufficient geographical size for environmental factors to come to bear. In particular, although the propagation/transmission of features over space is different in genealogical

language groups, henceforth called (linguistic) lineages, they can still be modeled in an abstract sense as clustered distributions of linguistic isoglosses forming large geographical areas. Two cases in the macro-areal profile of Africa discussed in §2.1 can serve as examples, because they are constituted by languages which are remotely or even closely related genealogically: the Sahara spread zone only comprises Afro-Asiatic languages and the Bantu spread zone is formed by a yet younger and closely knit family.

Accordingly, Güldemann (2010: 582) hypothesized that linguistic lineages, too, may have a latitudinal rather than longitudinal axis orientation with growing geospatial size. This possible effect of the latitude spread potential can be tested more easily. In Güldemann and Hammarström (forthcoming), we used various data sets to test this hypothesis, because there is still no agreement in the linguistic community as to the exact genealogical composition of the world's languages. These are the Glottolog 2 (Nordhoff et al. 2013, henceforth G2) and the Ethnologue (Lewis et al. 2013, henceforth E17).[3] We also tested the hypothesis on first-order subfamilies of the two primary data sets.

The way to determine the axis ratio of a lineage is essentially the same as that used for large-scale feature aggregations treated in §2.2 above. We used the geographical positions (centerpoints) of languages as given in E17 or in individual sources for languages not listed there. A lineage's axis ratio is the ratio between the East-West expansion d_{ew} (as the distance in km between the eastern and western endpoint languages of a lineage) and the North-South expansion d_{ns} as the distance in km between the northern and southern endpoint languages of the lineage. A lineage's geospatial size for the purpose of the present analysis is simply determined by multiplying d_{ew} and d_{ns}.

When plotting axis ratio against geospatial size, the above hypothesis is confirmed. The geometric mean of the axis ratio of all lineages is commonly close to 1, that is, neutral with respect to a latitudinal or longitudinal shape. This is expected because many lineages are small and on a small geographical scale environmental factors should not make a discernible impact. However, from a certain size on, the expected linear relationship between the two dimensions emerges. Thus, for the G2 data set, as shown in Figure 3, taking only the 50 largest lineages linear regression gives a modest ($r \approx 0.39$) but significant ($p < .01$) trend, and taking only the 10 largest ones gives a stronger ($r \approx 0.89$) also significant ($p < 0.1$) relationship.

---

[3] Since the first source uses throughout classification criteria that are consistently oriented towards the standards of orthodox comparative methodology (cf. Campbell and Poser 2008) and gives a brief justification within this frame with pointers to more explicit argumentation, we consider it a more coherent data set on independent lineages.

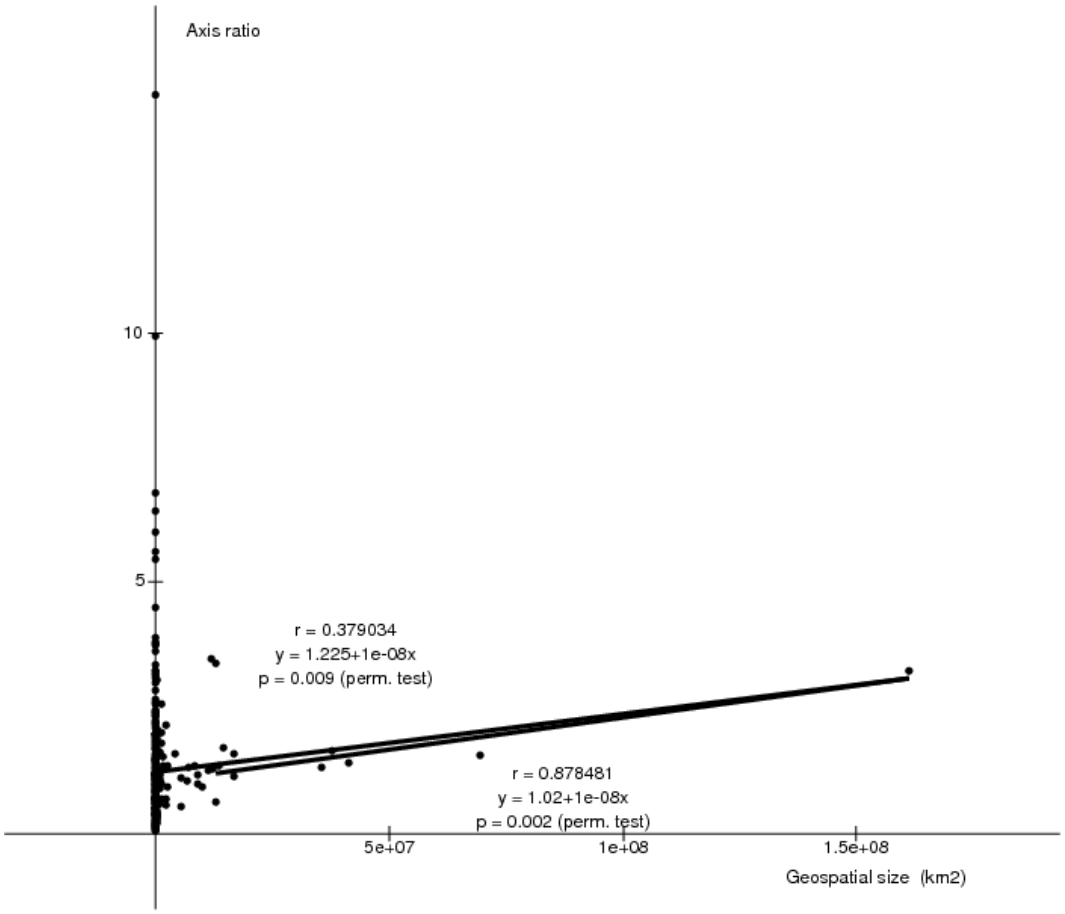


Figure 3: Size-axis ratio relationship for independent linages according to G2 with regression lines fit to the 50/10 largest ones.

This picture is largely replicated in the analysis of the E17 data set and the first-order subfamilies of both G2 and E17. All results are summarized in Table 4.

Data set	Geometric mean of axis ratio	Linear regression for 50 largest lineages	P	Linear regression for 10 largest lineages	P
G2	1.019	$r \approx 0.39$	$<.01$	$r \approx 0.89$	$<.01$
E17	1.144	$r \approx 0.46$	$<.01$	$r \approx 0.95$	$<.01$
G2 subfamilies	1.008	$r \approx 0.13$	$\approx.12$	$r \approx 0.99$	$<.05$
E17 subfamilies	1.084	$r \approx 0.18$	$\approx.06$	$r \approx 0.98$	$<.05$

Table 4: Relationship of axis-ratio and geospatial size across four data sets

Evidently, not all large lineages behave according to their geospatial size, because other more accidental contingencies like local geographical factors (e.g., geophysical barriers), particular histories that outplay any geographical constraints, etc. counteract the latitude spread potential. Such deviations should, however, cancel out, i.e., for every size range the lineages with a too high axis ratio should be matched by lineages with a too low axis ratio. We have therefore studied systematically the mean axis ratio according to variable size thresholds, improving on the arbitrary choice of top 10/50 largest lineages in

the expectation to find a smoothly rising curve rather than a randomly fluctuating one. Figure 4 shows for the G2 data set the mean axis ratio of all lineages larger or equal to the size of the nth largest lineage, as n ranges from all lineages down to one, confirming the expected trend. Again, this also holds overall for the other three data sets. While this tendency is weak for the class of small and medium-size lineages, it nevertheless shows up when considering means.

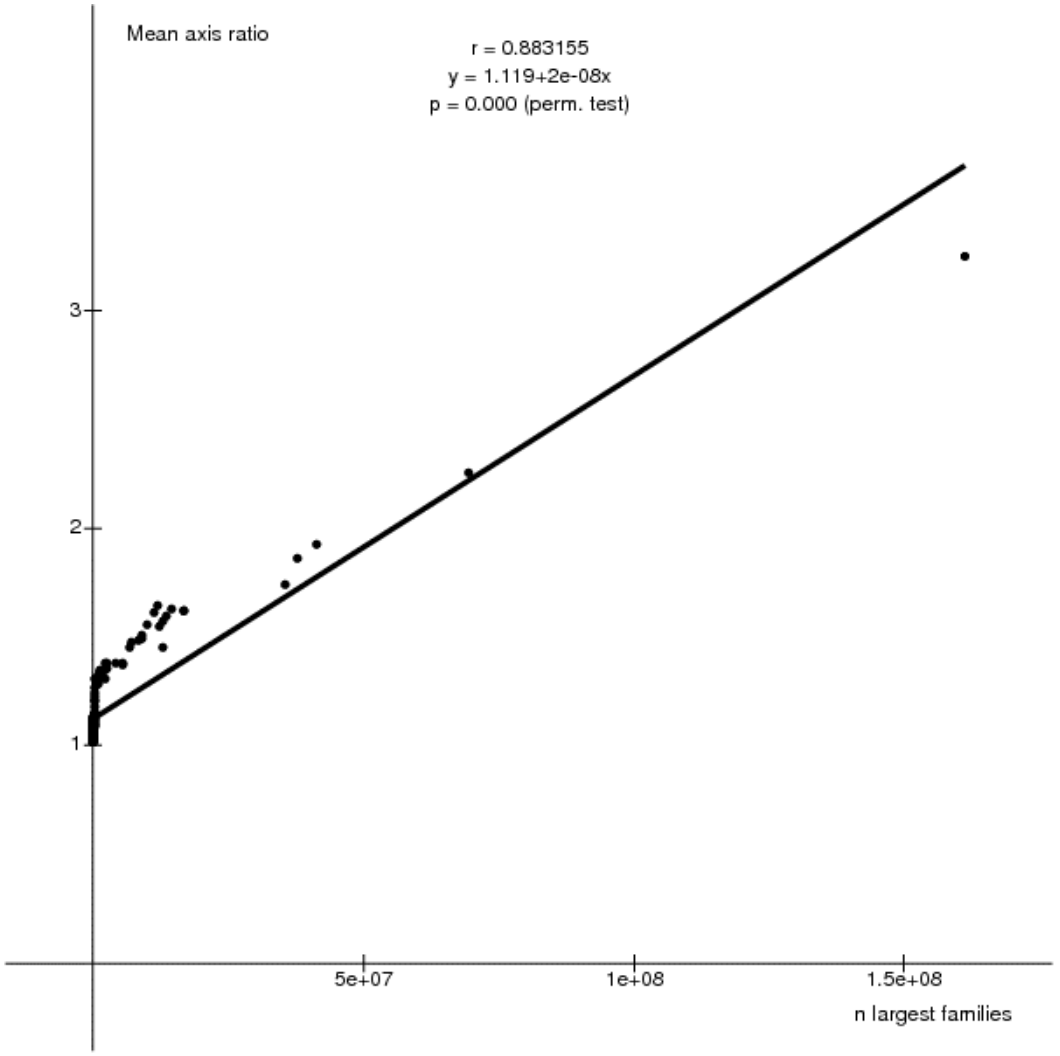


Figure 4: Size-threshold vs. mean axis ratio relationship for independent lineages according to G2 with regression lines

When evaluating our conclusion that there is a trend of linguistic lineages towards a latitudinal shape which gains in prominence with increasing geospatial size, it is also important to take the three following points into account. First, the latitude bias correlates only with geospatial lineage size, in line with our hypothesis. The observed trend disappears or becomes weaker as soon as lineage size is determined by other criteria, e.g., number of languages.

Second, we do not expect, and in fact do not find, that the trend necessarily holds for all (sub)continental areas separately, which elsewhere is a good criterion for large-scale typological investigations (cf. Dryer 1989). This is because other potentially opposing geographical factors can dampen or annihilate the latitude spread potential on a more local scale. Factors at issue in this respect seem to be particularly pronounced in Central and South America, in that the land masses have a strong north-south orientation (cf. Diamond 1999), the Andes have a strong north-south axis which may tend to steer the movement of (features associated with) human populations along longitudes, and even climate zones may be more longitudinal in South America (cf. Ricklefs 2001). Indeed, if testing the hypothesis on this geospatially large zone alone, no latitude bias of large lineages is found. With this background, it is important to acknowledge that the latitude spread potential is strong enough to show up on a world-wide scale.

Third and finally, looking at the overall shape of the earth's landmasses inhabited by humans one impressionistically observes that they are overall latitudinal and might thus be tempted to assume that this fact induces the global trend. We have designed a model of random growth of language families in order to control for this possible geographical contingency for which the reader is referred to Güldemann and Hammarström (submitted). The result is that landmass shape alone can indeed be held responsible for a certain amount of latitude bias of lineages but importantly is not sufficient to account for the degree of bias found in the real world. We can thus conclude overall that our initial hypothesis remains valid that the latitude spread potential is a likely factor for the global trend of lineages to have a latitudinal rather than longitudinal axis orientation with growing geospatial size.

2.3.2 The longitude spread constraint and intra-lineage diversity

The inverse of the latitude spread potential is the longitude spread constraint, namely a propensity towards changing environmental conditions that relatively speaking hamper rather than facilitate long-distance spread of populations and their features. Clearly, north-south movements are possible and amply attested historically. However, in these cases a geographical axis effect can be expected to turn up in a different way: generally speaking, populations more often confronting different kinds of barriers when moving in longitudinal direction are more prone to local adaptation and thus change. To mention just one scenario directly relevant for linguistics, greater environmental challenges encountered by a colonizing group in a new area are more likely to necessitate intensive contact with autochthonous populations, also leading potentially to considerable change in its linguistic profile, as hypothesized by Güldemann (2010: 580-2). Permanent repetition of this phenomenon over long time spans and sufficiently large space tends to steer intra-lineage linguistic diversity to be higher along north-south axes and lower along east-west axes.

An intuitively suggestive case in point seem to be pidgins and creoles with European lexifier languages which are predominantly distributed in the tropics and may structurally

even cluster together irrespective of their particular location (cf. Kortmann (2013) on Anglophone linguistic variation). Mainstream historical-comparative linguistics does not treat them as members of the relevant European lineage. However, one way to model their very emergence can be potentially related to the phenomenon entertained here: the European languages drastically changed in a certain geographical, historical and social setting, up to a point of arguably shifting away from their original genealogical alliance. One aspect, viz. widespread demographic inequality of the colonizers vis-à-vis the other populations (cf. Mufwene 2001, 2008), was at least partly steered by a drastically different environment - a challenge which the colonizers could not compensate by their strong and long-lasting socio-political and economic dominance.

One can also identify some relevant candidates of intra-family diversity from the African panorama dealt with in §2.1, e.g., Southern Cushitic as opposed to Cushitic in the Chad-Ethiopia area, or Northern Songhai as opposed to Songhai close to/in the Macro-Sudan belt. A particularly dramatic case seems to be the larger portion of Narrow Bantu languages in their own spread zone south of the Macro-Sudan belt, which is their place of origin and still hosts their closest relatives (cf. the discussion by Güldemann 2011).

The above observations are still purely impressionistic, though. For a first more systematic test regarding a possible axis bias of intra-lineage linguistic diversity we investigated the data on transitive sentence word order (see §2.2) in more detail. We first identified all pairs of genealogically related languages according to G2; these numbered 713193. For the record, 62.7% of these pairs displayed the same transitive sentence order (which was virtually the inverse of the situation for pairs of unrelated languages, where 65.5% of the total differed). We then classified the related-languages pairs according to whether they were more distant from each other on a latitudinal or longitudinal axis (henceforth just “latitudinal” and “longitudinal” (language) pairs). In a next step we determined the proportion of language pairs that disagree in the linguistic feature at regular distance intervals of 50km in both sets of pairs. For example, a pair of languages having a latitude distance of 75km and a longitude distance of 45km is put in the set of longitudinal language pairs within the distance interval of 50-100km. In this interval, there happened to be 4537 longitudinal pairs and 4813 latitudinal pairs with little difference between them in terms of feature change, namely a proportion of 9.7% vs. 9.3%, respectively.

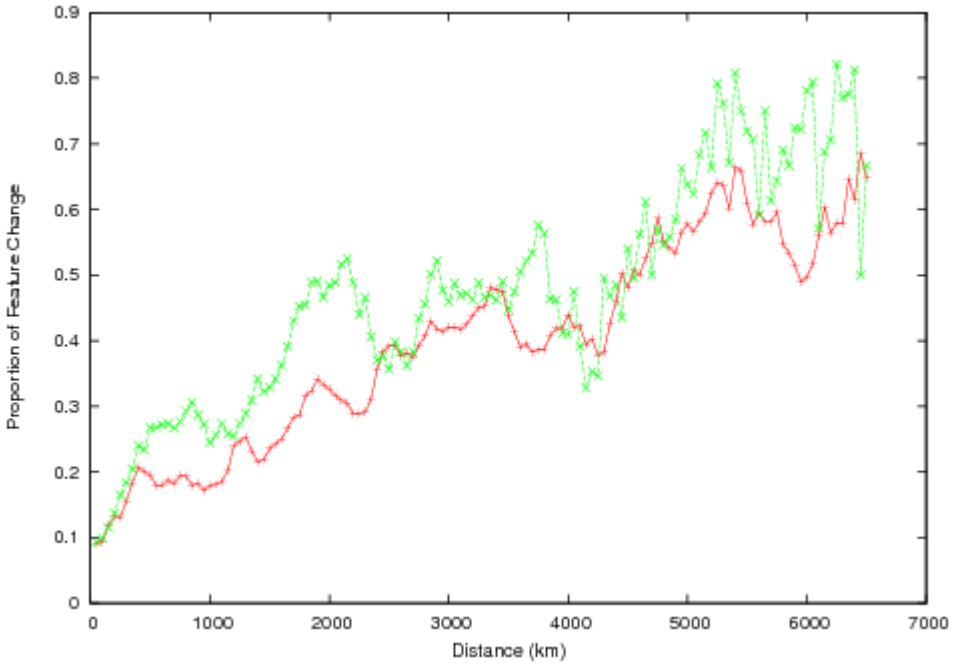


Figure 5: Proportion of change of transitive sentence order in latitudinal (red line) and longitudinal (green line) language pairs according to distance intervals

The overall results are shown in Figure 5. Up to 400kms latitudinal and longitudinal language pairs hardly differ with respect to feature change. From then on the longitudinal pairs consistently and by a margin show more feature disagreement than the latitudinal pairs at the same distance level, with an exception at a point near 1300kms. From ca. 2500kms on there is again no clear pattern of difference between the two sets of language pairs, although there are tops and dips that may be due to individual language families. After 5500kms both the latitudinal and longitudinal pairs reach an essentially random behavior of feature change found also with unrelated languages.

It is important to reflect on what is to be expected realistically under the above longitude constraint hypothesis. On short distances where environmental factors overall do not differ considerably latitudinal and longitudinal language pairs can be assumed to behave similarly. In the same vain language pairs that are geographically very far apart tend to be separated by a long different history so that any common heritage is likely to have vanished regardless of the pairs' axis configuration. In other words, pairs of related languages should behave like random language pairs at some distance level. So it is at a certain distance interval where one can expect that longitudinal language pairs display a feature change more often than latitudinal ones. The finding of our admittedly very restricted and preliminary test that this is indeed the case between 400 and 3000km is compatible with that expectation and indicates that our hypothesis deserves further exploration.

3 Conclusions

We have argued above that different kinds of linguistic data are compatible with the hypothesis that human migration or exchange and the accompanying spread of linguistic features over large distances are facilitated along latitude axes (tested for linguistic lineages and contact areas) and hampered along longitude axes (tested for contact areas). Thus, both latitude spread potential and latitude spread constraint influence the expansion, sedimentation, and retention of linguistic features over long time spans. This adds evidence to Diamond's idea that bio-geographical factors contribute to determining human history.

However, all tests reported on above are only first steps which need and can be refined in order to replicate our empirically still restricted findings. With respect to language families, the above test should be repeated with a genealogical classification of the world's languages that is more and more consensual by being permanently and collectively updated using strict and consistent criteria. For linguistic contact areas, both the range of areas and linguistic features need to be extended. Testing family-internal diversity and its geographical patterning in the future requires even more data breadth and methodological sophistication. First of all, a larger feature set using data bases of the magnitude of WALS but with a higher density is necessary, as evidenced also in the approaches by Cysouw and Comrie (2007) and Cysouw (2013). Moreover, not only feature diversity but also intra-family phylogenetic structure should be investigated with respect to geographical axis effects.

Despite the largely preliminary nature of the results presented above, this study may help to reinstate geography (beyond the universally acknowledged aspect of distance) as an important factor for the dynamics of areal linguistics on the macro-level. However, direct "mechanic" correlations between geography and linguistic distributions cannot be expected. As is evident, e.g., from the "minimal size-factor" concerning geographical axis effects, the patterns emerging from investigations of this type need instead to be embedded realistically in the relevant discipline and the nature of its phenomena and data.

Moreover, trying to explain large-scale feature aggregations with the help of geography can only provide one piece in a very complex puzzle. Any synchronic pattern is the result of a complex and long-term interplay of many different factors; these may conflict and their significance shift from historical period to historical period. Even conflicting pressures by two different factors may be stable (as suggested by the northern Great Rift Valley in Africa potentially being an area of "increased turbulence" without much feature sedimentation).

We also do not assume any form of extreme environmental determinism in the sense that geographical factors could not be outranked by other factors determining human behavior and history. Geographical factors have actually lost some of their impact on human population dynamics along the historical trajectory of our species in favor of other, notably socio-cultural factors.

The basic idea behind our discussion has been pronounced most prominently outside linguistics and indeed should be independent of it. One should therefore expect geographical axis effects in other non-linguistic aspects of human populations, too, such as cultural anthropology, physical and molecular anthropology, archeology, etc. In some of these fields, some attempts to test the basic idea have indeed been made, e.g., by Turchin, Adams and Hall (2006) on the east-west orientation of large historical empires and modern states, and by Laitin, Moortgat and Robinson (2012) on increased retention of cultural diversity on a north-south axis. These studies turned out to be far more complex in lacking a sufficiently large data base and/or having to control for a number of additional factors. Thus, linguistic features may prove to be particularly suitable for testing hypotheses like Diamond's.

References

Auer, Peter et al. (eds.). 2013. Space in language and linguistics: geographical, interactional, and cognitive perspectives. Berlin: Mouton de Gruyter.

Bickel, Balthasar and Johanna Nichols. 2012. Oceania, the Pacific Rim, and the theory of linguistic areas. In Antic, Zhenya et al. (eds.), Proceedings of the 32nd Annual meeting of the Berkeley Linguistic Society, February 10-12, 2006. Berkeley: Berkeley Linguistics Society 32, 3-15.

Bisang, Walter. 2006a. Contact-induced convergence: typology and areality. In Brown, Keith (ed.), Encyclopedia of Language and Linguistics, vol. 3. Oxford: Elsevier, 88-101.

Bisang, Walter. 2006b. Linguistic areas, language contact and typology: some implications from the case of Ethiopia as a linguistic area. In Matras, McMahon and Vincent (eds.), 75-98.

Bostoen, Koen A. G., Rebecca Grollemund and Joseph K. Muluwa. 2013. Climate-induced vegetation dynamics and the Bantu expansion: evidence from Bantu names for pioneer trees (Elaeis guineensis, Canarium schweinfurthii, and Musanga cecropioides). Comptes Rendus Geoscience 345: 336–349.

Campbell, Lyle. 2006. Areal linguistics: a closer scrutiny. In Matras, McMahon and Vincent (eds.), 1-31.

Chang, Will, Lev Michael and Tammie Stark. forthcoming. A statistical test for language contact. Language Dynamics and Change.

Clements, Nick and Annie Rialland. 2008. Africa as a phonological area. In Heine and Nurse (eds.), 36-87.

Collard, Ian F. and Robert A. Foley. 2002. Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? Evolutionary Ecology Research 4: 371-383.

Comrie, Bernard. 2007. Areal typology of Mainland Southeast Asia: what we learn from the WALS maps. Manusya: Journal of Humanities 13: 18-47.

Cullen, Ben S. 2000. Contagious ideas: on evolution, culture, archaeology, and cultural virus theory. Oxford: Oxbow.

Cysouw, Michael. 2013. Disentangling geography from genealogy. In Auer et al. (eds.), 21-37.

Cysouw, Michael and Bernard Comrie. 2009. How varied typologically are the languages of Africa. In Botha, Rudie and Chris Knight (eds.), The cradle of language. Oxford University Press, 189-203.

Dahl, Östen. 2001. Principles of areal typology. In Haspelmath, Martin et al. (eds.), Language typology and language universals: an international handbook, 2 vols. Berlin/ New York: Walter de Gruyter, vol. 2: 1456-1470.

Daumé, Hal, III. 2009. Non-parametric Bayesian areal linguistics. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09), 593–601. Morristown, NJ, USA: Association for Computational Linguistics.

Diamond, Jared M. 1999 [1997]. Guns, germs, and steel: the fates of human societies. New York/ London: W. W. Norton.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. Studies in Language 13,2: 257-292.

Ember, C. R. and M. Ember. 2007. Climate, econiche, and sexuality: influences on sonority in language. American Anthropologist 109: 180–185.

Enfield, Nick J. 2003. Linguistic epidemiology: semantics and grammar of language contact in mainland Southeast Asia. London: Routledge Curzon.

Enfield, Nick J. 2005. Areal linguistics and mainland Southeast Asia. Annual Review of Anthropology 34: 181-206.

Enfield, Nick J. 2008. Transmission biases in linguistic epidemiology. Journal of Language Contact 2: 295-306.

Enfield, Nick J. (ed.). 2011. Dynamics of human diversity: the case of mainland Southeast Asia. Canberra: Australian National University.

Everett, C. 2013. Evidence for direct geographic influences on linguistic sounds: the case of ejectives. PLoS ONE 8,6.

Fought, John G. et al. 2004. Sonority and climate in a world sample of languages. Cross-cultural Research 38: 27-51.

Greenberg, Joseph H. 1959. Africa as a linguistic area. In Bascom, William R. and Melville J. Herskovitz (eds.), Continuity and change in African cultures. Chicago University Press, 15-27.

Greenberg, Joseph H. 1983. Some areal characteristics of African languages. In Dihoff, Ivan R. (ed.), Current approaches to African linguistics 1. Dordrecht/ Cinnaminson: Foris, 3-22.

Güldemann, Tom. 1998. The Kalahari Basin as an object of areal typology - a first approach. In Schladt, Mathias (ed.), Language, identity, and conceptualization among the Khoisan. Köln: Rüdiger Köppe, 137-169.

Güldemann, Tom. 1999. Head-initial meets head-final: nominal suffixes in eastern and southern Bantu from a historical perspective. Studies in African Linguistics 28,1: 49-91.

Güldemann, Tom. 2003. Logophoricity in Africa: an attempt to explain and evaluate the significance of its modern distribution. Sprachtypologie und Universalienforschung 56,4: 366-387.

Güldemann, Tom. 2005. Complex predicates based on generic auxiliaries as an areal feature in Northeast Africa. In Voeltz, F. K. Erhard (ed.), Studies in African linguistic typology. Amsterdam: John Benjamins, 131-154.

Güldemann, Tom. 2007. Linguistic areas without evidence of contact. Paper presented at the International Symposium "Language Contact and the Dynamics of Language: Theory and Implications", Leipzig, May 10-13, 2007.

Güldemann, Tom. 2008. The Macro-Sudan belt: towards identifying a linguistic area in northern sub-Saharan Africa. In Heine and Nurse (eds.), 151-185.

Güldemann, Tom. 2010. Sprachraum and geography: linguistic macro-areas in Africa. In Lameli, Alfred, Roland Kehrein and Stefan Rabanus (eds.), Language and space: an international handbook of linguistic variation, volume 2: language mapping. Berlin: Mouton de Gruyter, 561-585, Maps 2901-2914.

Güldemann, Tom. 2011. Proto-Bantu and Proto-Niger-Congo: macro-areal typology and linguistic reconstruction. In Hieda, Osamu, Christa König and Hirosi Nakagawa (eds.), Geographical typology and linguistic areas, with special reference to Africa. Amsterdam: John Benjamins, 109-141.

Güldemann, Tom. 2013. Typology. In Vossen, Rainer (ed.), The Khoesan languages. London: Routledge, 25-37.

Güldemann, Tom. forthcoming. “Khoisan” linguistic classification today. In Güldemann, Tom and Anne-Maria Fehn (eds.), Beyond ‘Khoisan’: historical relations in the Kalahari Basin. Amsterdam: John Benjamins.

Güldemann, Tom and Anne-Maria Fehn. forthcoming. The Kalahari Basin area as a “Sprachbund” before the Bantu expansion - an update. In Hickey, Raymond (ed.), The Cambridge handbook of areal linguistics. Cambridge University Press.

Güldemann, Tom and Harald Hammarström. submitted. Global profile of language families supports geographical-axis hypothesis.

Hammarström, Harald and Tom Güldemann. forthcoming. Quantifying geographical determinants of large-scale distributions of linguistic features. Language Dynamics and Change.

Haspelmath, Martin et al. (eds.). 2005. The world atlas of language structures. Oxford University Press.

Heine, Bernd. 1976. A typology of African languages based on the order of meaningful elements. Berlin: Dietrich Reiner.

Heine, Bernd and Derek Nurse (eds.). 2008. A linguistic geography of Africa. Cambridge University Press.

Idiatov, Dmitry. 2010. Clause-final negation as a Macro-Sudan areal feature. Paper presented at the International Conference "Syntax of the World’s Languages 4", Laboratoire Dynamique du Langage (DDL) Lyon, September 23-26, 2010.

http://webh01.ua.ac.be/dmitry.idiatov/talks/2010_SWL4_Idiatov.pdf

Kortmann, Bernd. 2013. How powerful is geography as an explanatory factor in morphosyntactic variation? Areal features in the Anglophone world. In Auer et al. (eds.), 165-194.

Laitin, David D., Joachim Moortgat and Amanda L. Robinson. 2012. Geographic axes and the persistence of cultural diversity. Poceedings of the National Academy of Sciences of the Unites States of America 109,26: 10263-10268.

Lewis, Paul M., Gary F. Simons and Charles D. Fennig. 2013. Ethnologue: Languages of the World. 17th edn. Dallas: SIL International.

Lobeck, Armin K. 1946. Physiographic diagram of Africa. New York: Columbia University, Geographical Press.

Lucas, Christopher, Madeleine Cule and Iain Mathieson. 2011. Beyond 'eyeballing': towards and objective assessment of areal distributions in WALS. Paper presented at the workshop Towards greater objectivity in historical linguistics, International Conference in Historical Linguistics, July, 2011.

Matras, Yaron, April McMahon and Nigel Vincent (eds.). 2006. Linguistic areas: convergence in historical and typological perspective. Hampshire: Palgrave Macmillan.

Mufwene, Salikoko S. 2001. The ecology of language. Cambridge University Press.

Mufwene, Salikoko S. 2008. Language evolution: contact, competition and change. London/ New York: Continuum International.

Muysken, Pieter. 2000. From linguistic areas to areal linguistics: a research proposal. In Gilbers, Dicky, John Nerbonne, and Jos Schaeken (eds.), Languages in contact. Amsterdam/ Atlanta: Rodopi, 263-275.

Muysken, Pieter. 2008. Introduction. In Muysken, Pieter (ed.), From linguistic areas to areal linguistics. Amsterdam: John Benjamins, 1-23.

Muysken, Pieter et al. forthcoming. Linguistic areas, bottom up or top down? The case of the Guaporé-Mamoré region. In Comrie, Bernard and Lucia Golluscio (eds.), Areal linguistics in South America. ???.

Nettle, Daniel. 1999. Linguistic diversity. Oxford University Press.

Nichols, Johanna. 1992. Linguistic diversity in space and time. Chicago: University of Chicago Press.

Nichols, Johanna. 1997. Modeling ancient population structures and movement in linguistics. Annual Review of Anthropology 26: 359-384.

Nichols, Johanna. 2013. The vertical archipelago: adding the third dimension to linguistic geography. In Auer et al. (eds.), 39-60.

Nordhoff, Sebastian et al. 2013. Glottolog 2.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org. Accessed on 2013-07-02.

Ricklefs, R. E. 2001. The economy of nature. New York: W. H. Freeman.

Stolz, Thomas. 2002. No Sprachbund beyond this line! On the age-old discussion of how to define a linguistic area. In Ramat Paolo and Thomas Stolz (eds.), Mediterranean languages: papers from the MEDTYP Workshop, Tirrenia, June 2000. Bochum: Brockmeyer, 259-281.

Stolz, Thomas. 2006. All or nothing. In Matras, McMahon and Vincent (eds.), 32-50.

Thomason, Sarah G. 2001. Language contact: an introduction. Edinburgh University Press.

Turchin, Peter, Jonathan M. Adams and Thomas D. Hall. 2006. East-west orientation of historical empires and modern states. Journal of World-Systems Research 12: 219-229.