Classification: Biological Sciences: Genetics; Social Sciences: Psychology

Linguistic *tone* is related to the population frequency of the adaptive haplogroups of two brain size genes, *ASPM* and *Microcephalin*

Dan Dediu and D. Robert Ladd

School of Philosophy, Psychology and Language Sciences, the University of Edinburgh,
14 Buccleuch Place, Edinburgh, EH8 9LN, UK

Corresponding author: D. Robert Ladd, e-mail: bob@ling.ed.ac.uk.

Manuscript information: 23 pages, 1 figure and 0 tables.

Word and character counts: 202 words in the abstract and 46,785 characters in the paper.

Author contributions: DRL gathered most language data and ensured their encoding. DD gathered the population and genetic data and performed the statistical analyses.

Abbreviations footnote: *ASPM-D*, *MCPH-D*: the derived (adaptive) haplogroups of the genes *ASPM* and *Microcephalin*; *ns*: non-significant.

Abstract: The correlations between inter-population genetic and linguistic diversities are mostly non-causal (spurious), being due to historical processes and geographical factors that shape them in similar ways. Studies of such correlations usually consider allele frequencies and linguistic groupings (dialects, languages, linguistic families or phyla), sometimes controlling for geographic, topographic or ecological factors. Here, we consider for the first time the relation between allele frequencies and linguistic typological features. Specifically, we focus on the derived haplogroups of the brain growth and development-related genes *ASPM* and *Microcephalin*, which show signs of natural selection and a marked geographic structure, and on linguistic *tone*, the use of voice pitch to convey lexical or grammatical distinctions. We hypothesize that there is a relationship between the population frequency of these two alleles and the presence of linguistic tone, and test this hypothesis relative to a large database (983 alleles and 26 linguistic features in 49 populations), showing that it is not due to the usual explanatory factors represented by geography and history. The relationship between genetic and linguistic diversity in this case may be causal: certain alleles can bias language acquisition or processing and thereby influence the trajectory of language change through iterated cultural transmission.

Introduction. Human populations are diverse both genetically and linguistically, through inter-population differences in allele frequencies (1-3) and in the variety of languages and dialects they speak (4). In general, any relationship between these two types of diversity merely reflects geography and past demographic processes, not genetic influence on language behaviour (1,2,5-8). It is indisputable that normal infants of any genetic makeup can learn the language(s) they are exposed to in the first years of life, so we can assume with considerable confidence that there are no "genes for Chinese".

Nevertheless, it is well accepted that there is widespread inter-individual variation in many aspects relevant for language (developmental delays, differences in second-language learning aptitude. discrimination between foreign speech sounds (9), recognition of words in noise (10), differences in short-term phonological memory correlated with different syntactic processing strategies (11)). It is also accepted that this variation can be partially attributed to genetic factors, most probably through a "many genes with small effects" model including both generalist and specialist genes (12-15). There are also heritable aspects of brain structure in general, and language-related areas in particular (16-21).

It is therefore likely that there are heritable differences of brain structure and function that affect language acquisition and usage. These differences may have no obvious behavioral consequences in the non-clinical population; under ordinary circumstances, all normal speakers and hearers perform "at ceiling" on many language-related tasks (10). Moreover, no one doubts that all normal children acquire the language of the community in which they are reared. Nevertheless, if differences in language and speech-related capacities are variable and heritable and if the genes involved have inter-population structure, it is likely that populations may differ subtly in some of these aspects, and that differences between populations could influence the way languages change through cultural evolution over time.

It is generally acknowledged (22) that the process of language acquisition plays a major role in historical language change: language acquirers construct a grammar based on the language they hear around them, but the constructed grammar is not necessarily identical to that of their models, and the cumulative effect of such small differences over generations leads to language change. It follows that cognitive biases in a population of acquirers could influence the direction of language change across generations. These biasing effects could result in linguistic differences between populations, producing non-spurious (causal) correlations between genetic and linguistic diversities. Computer simulations (23,24) support the idea that such biases could influence the structure of languages emerging over many generations of cultural change, and mathematical models (25) suggest that, under appropriate conditions, extremely small biases at the individual level can be amplified by this process of cultural transmission and become manifest at the population level.

Linguistic tone. We propose that the linguistic typology of *tone* is affected by such a bias. Human languages differ typologically in the way they use voice fundamental frequency (pitch). All languages use consonants and vowels to distinguish one word or grammatical category from another, but, in addition, so-called ‘tone languages’ (e.g. Chinese) use pitch for this purpose as well, while ‘non-tone languages’ (e.g. English) use pitch only at sentence level (to convey emphasis, emotion, etc.) (26). In tone languages, that is, pitch is organised into tone phonemes that are functionally comparable to consonant and vowel phonemes. Tone languages are the norm in sub-Saharan Africa and are very common in continental and insular southeast Asia. They are rare in the rest of Eurasia, North Africa and Australia. They are relatively common in Central America, the Caribbean and the Amazon basin, and occur sporadically elsewhere among the aboriginal languages of the Americas (27).

The vast majority of the world's languages are unambiguously either tonal or not (27) but a few languages (e.g. Japanese, Swedish/Norwegian, Basque) are typologically intermediate, and it is well established that languages can lose or acquire tone through ordinary historical change (28). More strikingly, there are cases showing that the difference between "tonal" and "non-tonal" languages can actually be quite subtle, such as the existence of closely related (even mutually intelligible) languages and dialects of which some are "tonal" and some are not. The best described such cases are Kammu in Laos (29) and various Alaskan Athabaskan languages (30). In both cases the phonological interpretation of pitch differences associated with obstruent voicing (Kammu) or coda glottalisation (Athabaskan) is ambiguous in a way that could drive language change: specifically, these differences might be perceived by an acquirer either as part of a system of contrastive tones, or as allophonically conditioned accompaniments of glottalized or voiced obstruent phonemes. If, as we propose, tone is affected by some form of acquisition or processing bias, we might expect that it would manifest itself in cases like these. Though the exact nature of the bias is currently unclear, it is plausible that it might involve a propensity to favor linguistic structures in which elements such as phonemes and morphemes are strictly linearly ordered rather than (as is the case with tone) simultaneous or formally unordered.

A recent series of studies conducted by Wong and colleagues seems to point to inter-individual differences in tone learning and associated neural correlates (31,32). Adult speakers of a non-tonal language (English) were presented with an artificial language learning task involving lexical tonal distinctions, and it was found that they tend to form two groups, referred to as "successful" and "less successful" learners. A later study by the same team[*], focusing on the relationship between the anatomy of the primary auditory cortex and linguistic tone learning, found that the "successful" learners showed greater volume of left, but not right, Heschl's Gyrus, especially for gray matter.

[*] Dr. Patrick CM Wong, Dept. of Communication Sciences and Disorders, Northwestern University, 2240 Campus Dr., Evanston, IL 60208, pwong@northwestern.edu, (847) 491-2416 (phone), (847) 491-2429 (fax), *personal communication*, November 2006. Paper under preparation.

While this correlation could be entirely due to environmental effects of previous experience, it could also point to a genetic component. Interestingly, there are suggestions in the literature concerning the heritability of musical pitch processing (33) and the genetics of absolute pitch (34) and, while the relationship between linguistic and musical/absolute pitch is by no means simple (35), these studies are certainly consistent with the proposal of a genetic bias affecting linguistic tone.

ASPM and Microcephalin. *ASPM* (*MCPH5*, 1q31) and *Microcephalin* (*MCPH1*, 8p23) are two genes involved in brain growth and development (36-38). Deleterious mutations of both *ASPM* and *Microcephalin* are involved in recessive primary microcephaly (38-40), together with at least other four loci identified to date (39,41). During embryogenesis, the neuroepithelial cells, found around the telencephalic ventricle (42), undergo two types of division: *symmetric*, producing two neuroepithelial cells, or *asymmetric*, producing a neuroepithelial cell and a neuronal precursor (43) which migrates towards its final position in the cortex (42). The type of cell division is dependent on the orientation of the mitotic spindle relative to the apical-basal axis (43). It has been suggested (44) that a change in the number of symmetric divisions will dramatically alter brain size, given that each such division potentially doubles the final number of neurons. Both *ASPM* and *Microcephalin* are involved in cell-cycle regulation (45-48) and their deleterious mutations impact on the number of such symmetric divisions. It has been suggested that *ASPM* insures the maintenance of the perpendicular position of the mitotic spindle in the neuroepithelial cells, a very difficult task given their extremely elongated shape (43), which cannot be correctly accomplished by the truncated proteins associated with the deleterious mutations. Moreover, a recent report suggests a putative ciliary function for *ASPM* (49), pointing to an influence on neuronal migration, mediated by cerebrospinal fluid flow. For *Microcephalin*, the mechanism seems to be represented by the failure of the truncated protein to protect the neuroepithelial cells against DNA repair defects, leading to excessive apoptosis (39).

For both genes, “derived” haplogroups have been identified (the G allele for the A44871G polymorphism for *ASPM*, and the C allele for the G37995C polymorphism for *Microcephalin*) (36,37). These will be denoted as *ASPM-D* and *MCPH-D*, respectively. Their ages are estimated at 5.8ky (95%CI: 0.5-14.1ky) and 37ky (95%CI: 14-60ky), respectively, both showing signs of positive selection and a marked geographic structure (36,37). *ASPM-D* reaches high frequencies in Central and Western Asia, Europe and North Africa, as well as in Papua-New Guinea (but there are reasons to suspect contamination, see Discussion) and very low frequencies in East Asia, Sub-Saharan Africa and the Americas (see map in 36). *MCPH-D* is very frequent in Asia, Europe and the Americas, moderately frequent in North and East Africa, South-East Asia and Oceania (see comment on Papua-New Guinea), and very rare in Central, Western and South Sub-Saharan Africa (see map in 37). Moreover, both genes show signs of accelerated evolution in the human lineage (~2 favourable mutations/my; 38). The claim that the distribution of *ASPM-D* and *MCPH-D* is the result of positive selection has recently been challenged (50), but arguably remains the best explanation (51).

The phenotypic effects of the derived haplogoups of *ASPM* and *Microcephalin* are not yet known, but arguably do not include gross phenotypic alterations: the derived haplogroups are apparently not involved in variations in intelligence (52), brain size (53), head circumference, general mental ability, social intelligence (54) or the incidence of schizophrenia (55). We propose that their effects involve subtle differences in the organization of the cerebral cortex, with cognitive consequences including linguistic biases in the processing and acquisition of linguistic tone. More specifically, based on the suggestions in (43), it is highly possible that *ASPM-D* alters the orientation of the mitotic spindle dependent on local conditions in the precursors of language areas, leading to the emergence of the suggested bias. Moreover, it is plausible that *MCPH-D* contributes to these by influencing the number of symmetric divisions. One could envisage a hypothetical scenario whereby the changes induced by *MCPH-D* are enhanced by *ASPM-D* through a modification of the precise maintenance of the orientation of the mitotic spindle during the development of specific language-related areas.

The hypothesis. These considerations led us to hypothesise a relationship between the distribution of tone languages and the geographical structure of *ASPM-D* and *MCPH-D*. Those areas of the world where the new alleles are relatively rare also tend to be the areas where tone languages are common. As previously discussed, the effects of *ASPM-D* and *MCPH-D* on brain structure and functioning remain largely hypothetical, but it is entirely plausible that they influence the cognitive capacities involved in processing phonological structures, and thereby lead to linguistic biases of the type suggested above.

In the present study, we performed statistical tests of this hypothesis on the basis of a large database comprising 983 alleles and 26 linguistic features collected for 49 world populations (Materials and Methods), controlling for geographical and historical factors. We considered linguistic features rather than linguistic groupings (dialects, languages, linguistic families or phyla), because our hypothesis concerns specifically the interaction between linguistic typological diversity and population genetic diversity. We found that, in general, the relationship between these two diversities is fully explained by geographical and historical factors, whereas, in the specific case of *tone*, *ASPM-D* and *MCPH-D*, there is an important and significant correlation between their distributions even after controlling for geography and history. Therefore, we propose that this relationship is causal, that is, the genetic structure of a population can exert an influence on the language(s) spoken by that population. Further experimental support is required, but these findings suggest a fundamental direction for future research targeted at understanding the complex relationship between genetic factors, cultural evolution and linguistic phenomena.

Results

In the following, we have systematically applied Holm's multiple comparisons correction (56) and the reported *p*-values are adjusted. All the statistical analyses used R (57).

The relationship between linguistic features and alleles. The first aspect of the hypothesis concerns the existence of a relationship between the linguistic feature of *tone* and the derived haplogroups of *ASPM* and *Microcephalin*. We tested this by comparing the strength of the relationships between *tone* and *ASPM-D* and between *tone* and *MCPH-D* with the distribution of the relationships between all 26 linguistic features and all 983 genetic markers in our database. Specifically, we computed the distribution of the resulting values of Pearson's correlation coefficients, *r*, and found it to be normal for all pairs of linguistic features (N = 325, mean = 0.012, sd = 0.274), all pairs of alleles (N = 482,653, mean = 0.024, sd = 0.225), and all pairs of linguistic features and alleles (N = 25,558, mean = -0.006, sd = 0.218). This shows that, in general, linguistic features do not correlate with alleles. Focusing on the distribution of Pearson's *r* for all pairs of linguistic features and alleles, we found that the correlations between *tone* and *ASPM-D* and between *tone* and *MCPH-D* are both highly significant (*tone* and *ASPM-D*: $r = -0.53$, $p = 9.63 \bullet 10^{-5}$; *tone* and *MCPH-D*: $r = -0.54$, $p = 7.22 \bullet 10^{-5}$) and their values are in the top 1.5% of the empirical distribution of correlations.

This shows that, taken individually, *tone* and *ASPM-D*, and *tone* and *MCPH-D* are highly significantly correlated and the strength of their relationship is greater than 98.5% of all the 25,558 correlations between linguistic features and alleles in our database.

The relationship between linguistic features and *pairs* of alleles. The second aspect of our hypothesis concerns the relationship between *tone* and **both** *ASPM-D* and *MCPH-D*, which we tested using a logistic regression approach (58). We computed the logistic regressions of all linguistic features (as the dependent variables - DVs) on all pairs of alleles (as the independent variables - IVs)

(N = 11,582,690[†]), and their distribution is heavily skewed towards poor fit, as expected. However, the logistic regression of the DV *tone* on the IVs *ASPM-D* and *MCPH-D* is both very good (Nagelkerke's $R^2 = 0.528$, 73% correct classification; Intercept: estimate = 4.478, std. error = 1.843, $p = 0.015$; *ASPM-D*: estimate = -7.170, std. error = 2.767, $p = 0.010$; *MCPH-D*: estimate = -4.952, std. error = 2.217, $p = 0.026$) and in the top 2.7% of the empirical distribution of the logistic regressions. We also tested the effects of the interaction between *ASPM-D* and *MCPH-D* on *tone* (58), by performing the logistic regression of the DV *tone* on the IVs *ASPM-D*, *MCPH-D* and *ASPM-D*MCPH-D*, but the interaction term is *ns* ($p = 0.224$) and the new model does not perform better ($\chi^2(1) = 1.848, p = 0.174$).

This shows that *tone* and the pair *ASPM-D*/*MCPH-D* are highly significantly related and the strength of their relationship is greater than 97.3% of all the 11,582,690 converged logistic regressions between linguistic features and pairs of alleles in our database.

Controlling for geographical and historical factors. In order to control for the effects of geography and shared linguistic history on our results, we compared geographic, genetic, typological linguistic and historical linguistic distances between all pairs of populations in the sample. The *land (geographic) distances* are represented by great circle distances for pairs of populations on the same continent, with intercontinental paths forced through specific connection points (Damascus for Africa/Eurasia and Bangkok for Melanesia/Eurasia). The *genetic distances* are represented by Nei's *D* (59). For any set of linguistic features, the *typological linguistic distance* represents a generalized Euclidean distance over the space of these linguistic features (Materials and Methods). The *historical linguistic distance* is based on the linguistic relatedness given by historical linguistic classifications, as follows (60): it is *1* if the populations speak the same language, *2* if they speak languages belonging to the same branch of a linguistic family, *3* if they speak languages from different branches

[†] This is the number of logistic regressions for which the algorithm converged.

of the same linguistic family, and *4* if they speak languages not demonstrably related. Historical linguistic judgments are based on the classification in (4) and exclude controversial items.

We studied the relationships between these distances using Mantel (partial) correlations (61): $r = 0.509$, $p < 0.001$ (geographic vs. genetic); $r = 0.283$, $p < 0.001$ (geographic vs. typological linguistic); $r = 0.162$, $p = 0.011$ (genetic vs. typological linguistic) and $r = 0.021$, $p = 0.407$ (genetic vs. typological linguistic, while controlling for geographic distances). In general, therefore, the (weak) correlations between genetic and typological linguistic diversities can be accounted for by geography, confirming that, generally, there is no direct influence of genes on language behavior (2,5). As we are referring to typological linguistic diversity rather than the historically-based linguistic diversity of Cavalli-Sforza and coworkers (1), our results of a general lack of correlation between linguistic and genetic diversities do not contradict their findings.

Individually, the Mantel correlation with geography for *tone* is $r = 0.169$, $p = 0.015$, for *ASPM-D*, $r = 0.074$, $p = 1.000$ (due to Holm's multiple comparisons correction; 56), and for *MCPH-D*, $r = 0.543$, $p < 0.001$. Each of *tone*, *ASPM-D* and *MCPH-D* have low but significant spatial autocorrelations (62): Moran's I (63) is 0.178, 0.164 and 0.121, and Geary's c (64) is 0.634, 0.438 and 0.718, respectively, $p < 0.001$ for all, suggesting that, potentially, geographical factors might explain the observed relationship. However, the (partial) Mantel correlation between *tone* and the pair *ASPM-D*/*MCPH-D* is $r = 0.333$, $p < 0.001$, and, when controlling for geography, it decreases only slightly and still remains highly significant, $r = 0.291$, $p = 0.003$, showing that geography is not a good explanation for our empirical findings.

Tone, *ASPM-D* and *MCPH-D* tend to be much more similar inside than across linguistic families (the linguistic and genetic distances between populations speaking languages of the same families are smaller than across families: random permutations test (65), $p < 0.001$), suggesting that the shared linguistic history might explain the observed relationship between them. However, when controlling for the historical linguistic distances, the partial Mantel correlation between *tone* and the pair *ASPM-D*/*MCPH-D* remains important and highly significant ($r = 0.271$, $p < 0.001$), showing that the relationship cannot be fully explained in this manner.

Moreover, when controlling simultaneously for geography and shared linguistic history, the second-order partial Mantel correlation between *tone* and the pair *ASPM-D*/*MCPH-D* actually increases slightly and is highly significant ($r = 0.283$, $p < 0.001$), suggesting not only that geographical factors and shared linguistic history do not explain the hypothesized relationship, but that the linguistic history represents a suppressor variable (58) on this relationship.

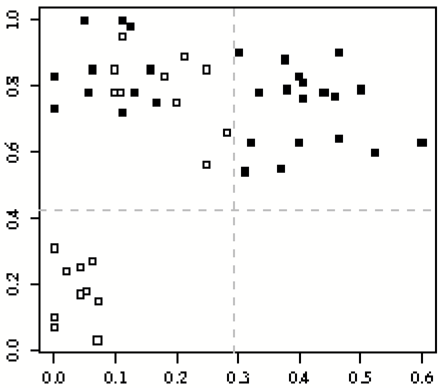

Fig. 1: Linguistic tone versus the population frequency of the adaptive haplogroups of ASPM and Microcephalin. The horizontal axis represents the frequency of ASPM-D, while the vertical axis represents the frequency of MCPH-D. Solid rectangles represent non-tonal languages and open rectangles tonal languages. Gray dashed lines correspond to 0.292 ASPM-D and 0.425 MCPH-D. See text for details.

Fig. 1 represents the distribution of linguistic *tone* as a function of the population frequency of *ASPM-D* and *MCPH-D*. Open rectangles stand for the *tonal* languages and their distribution corresponds to low frequencies of *ASPM-D* (lower than approximately 0.29), while solid rectangles stand for *non-tonal* languages and their distribution corresponds to high frequencies of *MCPH-D* (higher than approximately 0.42). Strikingly, in the bottom-left quadrant there are only tonal languages, in the top-right quadrant only non-tonal languages, while in the top-left quadrant there is an even distribution of tonal and non-tonal languages (10:11). There are no populations in our sample occupying the bottom-right quadrant. This figure illustrates the (probabilistic) predictions of our model concerning the tonality of a language given the frequency of *ASPM-D* and *MCPH-D* in that population. These predictions are corroborated by the 5 American populations not included in the analysis, which have low frequencies of *ASPM-D* and high frequencies of *MCPH-D*; as expected, their languages are both tonal and non-tonal. (We exclude the Papuan population from further consideration, as it seems likely to be unreliable due to contamination; see Materials and Methods). A very important test case for our model would be provided by Australia, since the Australian languages are non-tonal; however, obtaining reliable genetic samples seems very difficult.

Discussion

In this paper, we formulated and tested the hypothesis of a *non-spurious correlation* between linguistic *tone* and the derived haplogroups of two genes involved in brain growth and development, *ASPM* and *Microcephalin*. In so doing we have also introduced a novel methodological approach to studying the relationship between genetic and linguistic diversities. While we are well aware that a correlational approach cannot by itself prove causality, we have shown that our hypothesis is supported by the currently available data. Specifically, we have found that the negative correlation between *tone* and the population frequency of *ASPM-D* and *MCPH-D* cannot be explained by historical and geographical factors, thus strengthening the claim of a causal relationship between them. As noted in the introduction, we propose that the causal relation is mediated by a cognitive bias relevant to the processing and acquisition of tone.

We may summarize the structure of the proposed genetic influence on the distribution of linguistic tone in three necessary components or causal steps: from inter-individual genetic differences to differences in brain structure and function, from these to inter-individual differences in language-related capacities, and, finally, to typological differences between languages. The first component is represented by the proposed effects of *ASPM-D* and *MCPH-D* on brain structure and function, including the brain areas involved in linguistic tone. The second component involves inter-individual differences in the acquisition and/or the processing of tone, which are supported by several recent findings. The last component, probably the best supported to date, relies on the process of cultural transmission of language across generations, which can, in the right circumstances, amplify small individual biases to influence the trajectory of language change. We assume that any such bias is very small at the individual level and becomes manifest only at the population level through the process of cultural transmission. We also assume that the bias is probabilistic in nature and that many other factors, including language contact and history, also govern the process of language change and affect its outcome. Our findings therefore do not support any racist or deterministic interpretation. Finally, note that this bias could be either for or against tone, but the fact that non-tonality is associated with the derived haplogroups (Fig. 1) suggests that tone is phylogenetically older and that the bias favours non-tonality. The bias is presumably a selectively neutral byproduct of the two derived haplogroups, not connected to the selective pressures on them, as there is no evidence that tone itself confers any advantage or disadvantage on speakers. We cannot, of course, rule out the scenario whereby the natural selection detected for these haplogroups is partially due to their linguistic effects.

The correlation reported here represents the first plausible case in which differences in population genetic structure partially account for linguistic differences. This finding warrants future experimental work, which will help test and refine the hypothesis of a causal effect. The artificial language learning paradigm of Wong and colleagues (31) offers a solid framework for testing whether the existence of individual biases in the acquisition and processing of linguistic tone is influenced by the presence or absence of *ASPM-D* and *MCPH-D*. A study of the effects of these derived haplogroups on other language-related capacities, including phonological working memory or pitch tracking, is also warranted. Additionally, research is clearly needed on the phenotypic effects of these haplogroups on brain structure. Depending on the outcome of such experimental work, the results reported here could lead to a profound change in our understanding of the interactions between genetic diversity and our higher cognitive capacities, by bridging the gap between inter-individual and inter-population diversities. They also represent a solid foundation for gradual, accretionary models of language evolution and suggest a hitherto unsuspected mechanism driving language change.

Materials and Methods

The populations. The 49 populations used in this study were selected from the 59 populations in (36,37) based only on genetic and linguistic data availability. The Americas were too poorly sampled for their genetic and linguistic diversity, so that the 5 American populations have been excluded from the analysis, but used as a test case. *Orogen* is probably a misspelling of *Oroqen*. The populations have been identified geographically, linguistically and genetically using information from various sources (4,66-68; Maps Of World: http://www.mapsofworld.com/lat_long/index.html, accessed April 17, 2007). Due to systematically missing genetic information (see below), 4 African populations were eliminated (*Masai*, *Sandawe*, *Burunge* and *Zime*). Also, *Papuan* was eliminated due to its ambiguity and the high probability of contamination, suggested by its low genetic similarity to neighbors, but high to Europe. The *NAN Melanesian* (Non-Austronesian Melanesian) population is very poorly specified in (36,37), but it most probably represents (66,67) the *Naasioi* of Bougainville, Papua New Guinea. The 49 populations are[‡]: *Southeastern and Southwestern Bantu*, *San* (**naq**), *Mbuti Pygmy* (**efe**), *Turu* (**rim**), *Northeastern Bantu* (**kik**), *Biaka Pygmy* (**axk**), *Bakola Pygmy* (**gyi**), *Bamoun* (**bax**), *Yoruba* (**yor**), *Mandenka* (**mnk**), *Mozabite* (**mzb**), *Druze* (**apc**), *Palestinian* (**ajp**), *Bedouin* (**ayl**), *Hazara* (**haz**), *Balochi* (**bgp**), *Pathan* (**pst**), *Burusho* (**bsk**), *Makrani* (**bcc**), *Brahui* (**brh**), *Kalash* (**kls**), *Sindhi* (**snd**), *Hezhen* (**gld**), *Mongola* (**mvf**), *Daur* (**dta**), *Orogen* (**orh**), *Miaozu* (**hmy**), *Yizu*

[‡] Giving the 3 letter language codes (4). These linguistic attributions are not unique in some cases.

(**yif**), *Tujia* (**tji**), *Han* (**cmn**), *Xibo* (**sjo**), *Uygur* (**uig**), *Dai* (**tdd**), *Lahu* (**lhu**), *She* (**shx**), *Naxi* (**nbf**), *Tu* (**mjg**), *Cambodian* (**khm**), *Japanese* (**jpn**), *Yakut* (**sah**), *NAN Melanesian* (**nas**), *French Basque* (**eus**), *French (***fra**), *Sardinian* (**src**), *North Italian* (**vec**), *Tuscan* (**ita**), *Orcadian* (**sco**), *Russian* (**rus**) and *Adygei* (**ady**).

The genetic data. For each of the 49 populations, frequency and positional information was gathered about *ASPM-D* and *MCPH-D* (36,37), as well as about 133 alleles from the ALFRED database (66,67) and 1029 from the HDPG dataset (69), the only criterion being that frequency information is available for at least 44 of the 49 populations (the vast majority, except *ASPM-D* and *MCPH-D* are STRs). Positional information was obtained from the UniSTS Project (70), and for 50 no such information could be retrieved. Moreover, 124 pairs were duplicated between the two databases, and 9 were deleted as they introduced systematic missing data in sub-Saharan Africa. After these deletions, 981 alleles were retained.

Because genetic information is missing for most sub-Saharan populations, for the 5 populations speaking languages belonging to the Narrow Bantu branch of the Niger-Congo linguistic family (4) (*Southeastern and Southwestern Bantu*, *Turu, Northeastern Bantu*, *Bakola Pygmy* and *Bamoun),* the frequency information for the amalgamated "*Bantu speakers*" sample was used to replace the missing data. These 5 populations do not seem to be very different from the point of view of our genetic or linguistic data (paired samples t-tests between all pairs of these populations, separately for the linguistic and genetic data, are *ns*), and, moreover, they do not differ genetically from the "*Bantu speakers*" sample (paired t-tests are also *ns*). These results allow the amalgamation procedure, even if the demographic and linguistic histories of these 5 populations are very different (1,2). This procedure could introduce a bias towards those linguistic features uniform across the sampled Bantu languages and against those showing variation. To control for this, two artificial variants, *ASPM-D** and *MCPH-D*,* were created from *ASPM-D* and *MCPH-D*, by replacing their actual frequency values in the 5

Bantu populations with their averages. Systematic checks during all stages of the analysis suggest that this missing data handling procedure did not unduly distort the results.

The final database comprises 983 alleles, evenly distributed across the chromosomes. For each linguistic feature, the number of alleles correlating with it in the top 5% of the empirical distribution across the chromosomes does not deviate from the expected distribution (χ^2 tests *ns*), suggesting that there are no chromosomes tending to correlate better with the linguistic features.

The linguistic data. Of the 141 linguistic features in (71), 24 were retained. The criteria for retention were good coverage of the 49 populations and meaningful binary coding. Two new features (*Coda* and *OnsetClust*) were added. The 26 binary linguistic features, covering varied aspects of phonology and morpho-syntax, are: *ConsCat* (are there more than 25 consonants?), *VowelsCat* (are there more than 6 vowels?), *UvularC* (are there uvular consonants?), *GlotC* (are there glottalized consonants?), *VelarNasal* (are there velar nasals?), *FrontRdV* (are there front rounded vowels?), *Coda* (are codas allowed?), *OnsetClust* (are onset clusters allowed?), *WALSSylStr* (is syllable structure at least moderately complex as defined in (71)?), *Tone* (does the language have a tonal system?), *RareC* (does the language have any rare consonants?), *Affixation* (does the language use affixes?), *CaseAffixes* (are cases marked with affixes?), *NumClassifiers* (does the language have numeral classifiers?), *TenseAspect* (are there inflections marking tense-aspect?), *MorphImpv* (are there dedicated morphological categories for second person imperatives?), *SVWO* (what is the dominant Subject-Verb word order (if any)?), *OVWO* (what is the dominant Object-Verb word order (if any)?), *AdposNP* (what is the dominant order (if any) between adposition and noun phrase?), *GenNoun* (what is the dominant order (if any) between genitive and noun?), *AdjNoun* (what is the dominant order (if any) between adjective and noun?), *NumNoun* (what is the dominant order (if any) between numeral and noun?), *InterrPhr* (are 'WH' question words phrase-initial?), *Passive* (is there a passive construction?), *NomLoc* (are locational predication and nominal predication encoded the same way?)

and *ZeroCopula* (is omission of copula allowed?).

For each of the 49 populations, the values of these 26 linguistic features were collected. The attribution of values to these features was based, where possible, on published material (71-84), but we also gathered primary data by sending standardized questionnaires to specialists in several of the languages concerned (see Acknowledgments). In most instances, this attribution is straightforward, but in some it involves a certain degree of subjective judgment, while in some others the data is simply unavailable. Nevertheless, we are confident that most linguists would agree with the vast majority of our decisions.

The typological linguistic distance. For any set of linguistic features, f_i, and pair of populations, p_1 and p_2, the typological linguistic distance is defined as:

$$D_L(f_1,..,f_n; p_1, p_2; w_1,.., \mathrm{w}_n) = \sqrt{(\Sigma w_i(f_{1i}-f_{2i})^2)}$$

The *equal weighting scheme* considers all features equally important: $w_1 = ... = w_n = 1/n$. Let H_i be the informational entropy (85) of linguistic feature f_i; then the *direct proportion weighting scheme* considers more important those features which carry more information, $w_i = H_i / \Sigma H_i$, while the *inverse proportion weighting scheme* considers more important those features whose distribution is more skewed, $w_i = 1/(H_i \Sigma(1/H_i))$. These three weighting schemes intercorrelate extremely well (Mantel's $r = 0.996$, 0.978 and 0.959, respectively, $p < 0.001$), so that only the equal weighting scheme was used.

Acknowledgments

We thank B Connell, C Kutsch Lojenga, H Eaton, JA Edmondson, J Hurford, K Bostoen, L Ziwo, M Blackings, N Fabb, O Stegen, R Asher, R Ridouane, M Endl and J Roberts for help with language data, A Dima for help with statistics, and J Hurford, S Kirby, R McMahon, S Della Sala, T Bates and

P Wong for discussions and comments. We also thank three anonymous reviewers for their suggestions. DD was funded by an ORS Award and a Studentship from the College of Humanities and Social Science, University of Edinburgh.

1. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes* (Princeton University Press, Princeton, New Jersey).

2. Jobling MA, Hurles ME, Tyler-Smith C (2004) *Human Evolutionary Genetics* (Garland Science, New York).

3. Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) *Am J Hum Genet* 72:578-589.

4. Gordon RG, Jr. (2005) *Ethnologue: Languages of the World*, (SIL International, Dallas, Texas, ed. 15).

5. Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W. Armenteros M, Arroyo E, Barbujani G *et al.* (2000) *Am J Hum Genet* 67:1526-1543.

6. Bellwood P, Renfrew C eds (2002) *Examining the farming/language dispersal hypothesis* (McDonald Institute for Archaeological Research, Cambridge).

7. Bateman R, Goddard I, O'Grady R, Funk VA, Mooi R, Kress WJ, Cannell P (1990) *Curr Anthropol* 31:1-13.

8. MacEachern S (2000) *Curr Anthropol* 41:357-385.

9. Golestani N, Molko N, Dehaene S, LeBihan D, Pallier C (2007) *Cereb Cortex* 17:575-582.

10. Surprenant AM, Watson CS (2001) JASA 110:2985-2095.

11. Swets B, Desmet T, Hambrick DZ, Ferreira F (in press) *J Exp Psychol Gen* in press.

12. Stromswold K (2001) *Language* 77:647-723.

13. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco PA (2001) *Nature* 413:519-523.

14. Plomin R, Kovas Y (2005) *Psychol Bull* 131:592-617.

15. Bates TC, Luciano M, Castles A, Coltheart M, Wright MJ, Martin NG (2007) *Eur J Hum Genet* 15:194-203.

16. Wallace GL, Schmitt, EJ, Lenroot R, Viding E, Ordaz S, Rosenthal MA, Molloy EA, Clasen LS, Kendler KS, Neale MC, Giedd JN (2006) *J Child Psychol Psychiatry* 47:987-993.

17. Pennington BF, Filipek PA, Lefly D, Chhabildas N, Kennedy DN, Simon JH, Filley CM, Galaburda A, DeFries JC (2000) *J Cogn Neurosci* 12:223-232.

18. Wright IC, Sham P, Murray RM, Weinberger DR, Bullmore ET (2002) *Neuroimage* 17:256-271.

19. Bartley AJ, Jones DW, Weinberger DR (1997) *Brain* 120:257-269.

20. Thompson PM, Cannon TD, Narr KL, van Erp T, Poutanen VP, Huttunen M, Lönnqvist J, Standertskjöld-Nordenstam CG, Kaprio J, Khaledy M, Dail,R, Zoumalan CI, Toga AW (2001) *Nat Neurosci* 4:1253-1258.

21. Scamvougeras A, Kigar DL, Jones D, Weinberger DR, Witelson SF (2003) *Neurosci Lett* 338:91-94.

22. Trask RL (1996) *Historical linguistics* (Arnold, London).

23. Smith K (2004) *J Theor Biol* 228:127-142.

24. Nettle D (1999) *Lingua* 108:95-117.

25. Kirby S, Dowman M, Griffiths TL (*in press*) *PNAS.*

26. Cutler A, Dahan D, van Donselaar W (1997) *Lang Speech* 40:141-201.

27. Maddieson I (2005) in *The World Atlas of Language Structures*, eds Haspelmath M, Dryer MS, Gil D, Comrie B (Oxford University Press, Oxford).

28. Hyman LM (1978) in *Tone: A linguistic survey*, ed. Fromkin VA (Academic Press, London), pp. 257-269.

29. Svantesson J-O, House D (2006) *Phonology* 23:309-333.

30. Krauss ME (2005) in *Athabaskan Prosody*, eds Hargus S, Rice K (Benjamins, Amsterdam), pp. 51-137.

31. Wong PCM, Perrachione TK (in press) *Appl Psycholing* in press.

32. Wong PCM, Perrachione TK, Parrish TB (in press) *Hum Brain Mapp* in press.

33. Drayna D, Manichaikul A, de Lange M, Snieder H, Spector T (2001) *Science* 291:1969-1972.

34. Baharloo S, Service SK, Risch N, Gitschier J, Freimer NB (2000) *Am J Hum Genet* 67:755-758.

35. Patel A (in press). *Music, Language and the Brain* (Oxford University Press).

36. Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, Tishkoff SA, Lahn BT (2005) *Science* 309:1720-1722.

37. Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, Vaez-Azizi LM, Tishkoff SA, Hudson RR, Lahn BT (2005) *Science* 309:1717-1720.

38. Gilbert SL, Dobyns WB, Lahn BT (2005) *Nat Rev Genet* 6:581-590.

39. Cox J, Jackson AP, Bond J, Woods CG (2006) *Trends Mol Med* 12:358-366.

40. Woods CG (2004) *Curr Opin Neurobiol* 14:112-117.

41. Woods CG, Bond J, Enard W (2005) *Am J Hum Genet* 76:717-728.

42. Tang BL (2006) *Biochem Biophys Res Commun* 345:911-916.

43. Fish JL, Kosodo Y, Enard W, Pääbo S, Huttner WB (2006) *PNAS* 103:10438-10443.

44. Caviness VS Jr., Takahashi T, Nowakowski RS (1995) *Trends Neurosci* 18:379-383.

45. Trimborn M, Bell SM, Felix C, Rashid Y, Jafri H, Griffiths PD, Neumann LM, Krebs A, Reis A, Sperling K, Neitzel H, Jackson AP (2004) *Am J Hum Genet* 75:261-266.

46. Bond J, Woods CG (2006), *Curr Opin Cell Biol* 18:95-101.

47. Zhong X, Liu L, Zhao A, Pfeifer GP, Xu X (2005) *Cell Cycle* 4:1227-1229.

48. Zhong X, Pfeifer GP, Xu X (2006) *Cell Cycle* 5:457-458.

49. Ponting CP (2006) *Bioinformatics* 22:1031-1035.

50. Currat M, Excoffier L, Maddison W, Otto SP, Ray N, Whitlock MC, Yeaman S (2006) *Science* 313:172.

51. Mekel-Bobrov N, Evans PD, Gilbert SL, Vallender EJ, Hudson RR, Lahn BT (2006) *Science* 313:172b.

52. Mekel-Bobrov N, Posthuma D, Gilbert SL, Lind P, Gosso MF, Luciano M, Harris SE, Bates TC, Polderman TJC, Whalley LJ, Fox H, Starr JM, Evans PD, Montgomery GW, Fernandes C, Heutink P, Martin NG, Boomsma DI, Deary IJ, Wright MJ, de Geus EJC, Lahn BT (2007) *Hum*

Mol Genet Advance Access Jan. 12, 2007; doi:10.1093/hmg/ddl487.

53. Woods RP, Freimer NB, De Young JA, Fears SC, Sicotte NL, Service SK, Valentino DJ, Toga AW, Mazziotta JC (2006) *Hum Mol Genet* 15:2025-2029.

54. Rushton JP, Vernon PA, Bons TA (2007) *Biol Lett* 3:157-160.

55. Rivero O, Sanjuán J, Moltó M-D, Aguilar E-J, Gonzalez J-C, de Frutos R, Nájera C (2006) *Schizophr Res* 84:427-429.

56. Holm S (1979) *Scand J Stat* 6:65-70.

57. R Development Core Team (2006) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna). Online http://www.R-project.org.

58. Tabachnick BG, Fidell LS (2001), *Using multivariate statistics* (4th edition, Allyn & Bacon, Needham Heights, MA).

59. Nei M (1972) *Am Nat* 106:283-292.

60. Nettle D, Harriss L (2003) *Hum Biol* 75:331-344.

61. Mantel N (1967) *Cancer Res* 27:209-220.

62. Fortin M-J, Dale M (2005) *Spatial analysis: A guide for ecologists* (Cambridge University Press, Cambridge, UK).

63. Moran P (1948) *J Roy Stat Soc B* 10:243-289.

64. Geary RC (1954) *The Incorporated Statistician* 5:115-145.

65. Edgington ES (1995) Randomization tests, 3^{rd} Ed (Marcel Dekker, NY).

66. Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK (2002) *Am J Phys Anthropol* 119:77-83.

67. Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, Yeh CC, Miller PL, Kidd KK (2003) *Nucleic Acids Res* 31:270-271.

68 CIA The World Factbook (2007). Available at https://www.cia.gov/cia/publications/factbook/index.html. Accessed April 17, 2007.

69. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) *Science* 298:2381-2385.

70. The UniSTS Project. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unists. Accessed April 17, 2007.

71. Haspelmath M, Dryer MS, Gil D, Comrie B eds (2005) *The World Atlas of Language Structures* (Oxford University Press, Oxford).

72. Campbell GL (2000) *Compendium of the World's Languages* (Routledge, London, ed. 2, vols. 1 & 2).

73. Tucker AN, Mpaayei JTO (1955) *A Maasai Grammar* (Longmans Green, London).

74. Mugane JM (1997) *A Paradigmatic Grammar of Gikuyu* (CLSI Publications, Stanford).

75. Guthrie M (1948) *The Classification of the Bantu Languages* (Oxford University Press).

76. Guthrie M (1953) *The Bantu Languages of Western Equatorial Africa* (Oxford University Press,).

77. Penchoen TG (1973) *Tamazight of the Ayt Ndhir* (Undena Publications, Los Angeles).

78. Lazard G (1992) *A Grammar of Contemporary Persian* (Mazda Publishers in association with Bibliotheca Persica, Costa Mesa CA).

79. Schmitt R, ed (1989) *Compendium Linguarum Iranicarum* (Dr. Ludwig Reichert Verlag, Wiesbaden).

80. Bashir EL (1991) *A Contrastive Analysis of Brahui and Urdu* (Academy for Educational Development, Washington DC).

81. Masica CP (1991) *The Indo-Aryan Languages* (Cambridge University Press).

82. Xi Z (1996) *PhD Thesis*, Dept. Of Linguistics, University of Toronto.

83. Mortensen D (2006) *Preliminaries to Mong Leng (Hmong Njua) Phonology* (Ms. Online http://ist-socrates.berkeley.edu/~dmort/mong_leng_phonology.pdf). Accessed April 17, 2007.

84. *Organised Phonology Data: Nasioi* [government spelling] (Naasioi [language spelling]) Language [NAS] Kieta – North Solomons Province (online

http://www.sil.org/pacific/png/pubs/0000268/Nasioi.pdf). Accessed April 17, 2007.

85. Shannon CE (1948) *The Bell System Technical Journal* 27:379–423, 623–656.