**Supplementary Materials:**

Materials and Methods

Figures S1-S8

Tables S1-S6 (in a separate text file)

**Supplementary Materials:**

Material and Methods

This paper tests the relationship between syllable structures and mitochondrial DNA haplogroups in Africa and Eurasia. The paper focuses on Africa and Eurasia, in part because of the availability of extensive genetic data for which there is matching linguistic data. It is also to constrain the focus of interactions between language families to within the last ten thousand years, during which it is unlikely that there has been pre-colonial linguistic interaction of African and Eurasian languages with languages in the Americas or Australia.

A consensus on language family classification was followed using Glottolog *(29*). Languages were assigned a geographical coordinate provided by the World Phonotactics database *(21)*, or the World Atlas of Language Structures (*27*) in the case of word order.

Data on syllable structure was taken from the World Phonotactics Database *(21)*. Three properties were coded for: the number of consonants allowed in the onset of a syllable, the number of consonants allowed in the coda, and the number of tones.

The coding in the database is based on native vocabulary and attempts to exclude loan-words. Some languages are coded as having consonant clusters despite this being restricted (such as only being able to add -*s*). Other potential issues of coding are the way that semi-vowels and glides are counted. For the sake of consistency, the coding in the database was taken as it was, without changing any of the coding, although it is possible that in some cases the coding could be disputed.

Data on haplogroups was taken from papers on mitochondrial DNA *(7-20*). This data is summarized in table S1. The populations were then compared using 252 haplogroups. These ranged from the most general to the most specific. The frequencies of each haplogroup were calculated using table S1 and Phylotree *(30)*, the phylogeny of mitochondrial DNA haplogroups which is summarized in Newick tree format in table S2 (dummy nodes are included in order to make it readable in the R ‘ape’ package, which does not read trees with singleton nodes)*.* If there is a specific haplogroup present in a population, then the haplogroups of which it is a sub-clade are also counted as being in that population. For example, if there is 20% haplogroup H1 and 10% haplogroup H2, then the total frequency of haplogroup H in the population is 30%. The resulting table of populations with frequencies for each of the 252 haplogroups is given in table S3, along with data on syllable structure.

Statistical analysis was carried out using R, in particular with the ‘lme4’ package for mixed effects models, ‘ade4’ for the Mantel tests, ‘mapdata’ for producing the maps and the ‘ape’ package for the phylogeny of mtDNA haplogroups. The method was to test all three syllable structure features against all 252 haplogroups. This may seem like multiple testing, but it is justified because it is not know in advance which haplogroups will have carried distinctive syllable structures. The method of correcting for multiple testing was to compile the p-values of all the tests performed into a histogram, and to see how many p-values were below 0.05. In purely random data, the number of p-values below 0.05 would not be any higher than p-values in other intervals. For example, the histogram in figure S1 is generated using a purely random arrangement of the coda data and then tested against the 252 haplogroups. The bar representing the number of p-values below 0.05 is not higher than the bars representing p-values in other intervals. By contrast, using the actual coda complexity data produces the histogram in figure S2, in which the number of p-values below 0.05 is clearly higher than the number of p-values in other intervals. The second greatest number of p-values is in the range 0.3<p<0.35. There are 77 correlations with p-values below 0.05 compared with 28 in the interval 0.3<p<0.35, suggesting that at least 49 correlations are genuine rather than merely due to multiple comparisons. Onset complexity and number of tones also showed this pattern (Figs. S3 and S4).

This method is only able to show that there are genuine correlations in the data, not which ones are genuine and which ones are spurious. The next step is to look at the strongest correlations, in particular any one which survives a Bonferroni correction for the number of tests done. One correlation which survives a Bonferroni correction for the 756 tests done is the correlation between coda complexity and haplogroup HV (p=3.8e-08).

A different method was to use a Mantel test of overall genetic distance and difference in syllable complexity. In this case genetic distance was calculated simply as the total sum of the differences in frequencies of each haplogroup, without taking into account number of mutations each haplogroup differed by (as an analysis of molecular variance would). The reason for this is that populations such as Khoisan populations would be very distant genetically from all other populations if mutational differences between Khoisan mtDNA haplogroups and other haplogroups were taken into account. The method is not trying to compare linguistic structures with the history of populations since the expansion of humans out of Africa, but within a time-frame closer to 10,000 years when already differentiated mtDNA haplogroups would have been moving between populations. Similarity between populations calculated by taking the Manhattan distance of the frequencies of all haplogroups would then be a proxy for how much different populations have been in contact (or whether they share common ancestry).

Mixed effects models were then tested to see if the correlations were purely due to non-independence of languages within known families. For each haplogroup coda complexity was tested in a mixed effects model with the haplogroup frequency and using language family as a random intercept. This mixed effects model was then compared to a model using just language family as a random intercept. These two models were then compared using the ANOVA function, and a p-value is given for whether the model with haplogroup frequency is significantly better than the model without. This was done for all 252 haplogroups, and the p-values compiled into a histogram, shown in figure S4. The number of p-values below 0.05 was again much higher than for other values, except for the interval between 0.95 and 1. This was also true when done for onset complexity and number of tones. The interval of 0.95<p<1 can be ignored because these p-values probably represent tests for which there is low data. This pattern was also found for onset complexity and number of tones, shown in figure S5. In all three cases the number of p-values below 0.05 is otherwise much higher than the p-values in other intervals.

Coda complexity correlates the strongest with haplogroup HV, plotted in figure S6. In order to investigate whether this was due to known language families, a mixed effects model was used and found to be significantly better than a model without haplogroup frequency as a fixed effect (χ2=5.39, p=0.02). Another way of controlling for the non-independence of related languages was to sample one language per family. In this case the correlation was significant 49% of the time, which is sufficiently above chance to show that there is a correlation after controlling for language family.

A Mantel test was done to see whether haplogroup HV and coda complexity correlated simply because they both correlate with geography. Distances between populations were calculated using the latitude and longitude of each population and the Haversine formula. The test was still significant controlling for both distance and language family (r=0.21, p=0.001), suggesting that the correlation is not simply due to the fact that both traits correlate with geography.

Word order features and phonemes were subsequently tested, using data from two independently coded databases *(26,27)*. Six different word orders were used: the order of subject and verb, object and verb, genitive and noun, adjective and noun, numeral and noun, demonstrative and noun, and relative clause and noun. These word orders features were split into three and binarised, namely whether they had placed the first element before the second, the second element before the first, or whether they allowed both orders; for instance, if a language places the numeral exclusively before the noun, then it will be coded as having 1 for numeral-noun order, 0 for noun-numeral order, and 0 for having both numeral-noun and noun-numeral orders. If a language allows both orderings, then it will be coded as having 0 for numeral-noun order, 0 for noun-numeral order, and 1 for having both. This resulted in 18 different features, which were tested against all 252 haplogroups again. It should be pointed out that these word order features are non-independent, both because having a value of 1 for one feature means that it has a value of 0 for the other two related features (e.g. having 1 for numeral-noun means that it has 0 for noun-numeral and 0 for both orderings), and because certain word orders may be mutually interdependent such as verb-object order and other word orders *(27)*. 1638 different phonemes were used from the PHOIBLE database, summarized in table S5. Since the languages in these databases were different from the languages analyzed for the syllable structure data, they were assigned populations with genetic data summarized in tables S4 and S6 respectively.

The same multiple testing method was used for word orders and phonemes. Word orders showed the same pattern as syllable structures in correlating with mtDNA haplogroups, although numeral-noun order was the only feature which showed this pattern individually (figure S7). Phonemes did not show this pattern initially, instead showing many correlations with high p-values (figure S8). This is probably due to low data for some phonemes and haplogroups, namely that many phonemes are extremely rare and occur in a single language in the sample, or none at all; a similar point holds for some haplogroups being found in a single population. One solution was to exclude phonemes and haplogroups which are found in less than 10% of the sample, which then showed that there are more phoneme-haplogroup correlations than expected by chance, as shown in the second histogram of figure S8. No inference was made about the spread of particular phonemes due to lack of space and due to the large number of phonemes tested, although there may be particular phoneme-haplogroup correlations which are especially significant. The main point was simply to demonstrate that mitochondrial DNA haplogroups correlate with linguistic features other than syllable structure as well.

In the extended African sample, a paper with information on haplogroup HV in different African populations was used *(24).* The paper does not give detailed linguistic information on the samples, instead giving the country that the samples are from. In order to assign a language to each sample, a median value was taken from each country; for example, the median number of consonants allowed in the onset in languages in Egypt is 1. The capital of the country was used to represent the location of the sample. Table S7 and table S8 summarize this data for syllable structure and numeral-noun order respectively. The correlation between haplogroup HV and numeral-noun ordering within African languages was tested but found to be not significant (p=0.3).

The frequency of HV in the populations in that paper was given, as was the frequency of haplogroup H. To get the actual frequency of macro-haplogroup HV, however, these two frequencies have to be added together, as in their paper HV means having HV but not having H. The paper also provided data on the frequency of haplogroup U, which was tested against the order of numeral and noun within African languages given that it correlated strongly with that feature in the whole sample. That correlation was highly significant (p=5.9e-06), although the correlation with haplogroup HV was not (p=0.3). This may show that the property of numeral-noun order may have a different historical spread from the property of consonant clusters, and hence correlating with different haplogroups.

A general point in this paper is that structural properties can be historically informative; this is in spite of the fact that they have a limited design space. Languages typically only have a choice of placing the verb before or after the object, for example, which means that two historically unconnected languages may share the same order of verb and object by chance. This means that structural properties cannot by themselves be taken as evidence of shared history; they should be shown first to have a historically informative pattern by testing their correlation against something whose history we have an understanding of, such as genetic variants. Many structural properties will thus not be very informative, such as most of the word order features tested here; but some will be such as numeral-noun ordering, especially if they can additionally be shown to be stable within accepted language families such as Indo-European, in which numeral-noun order is mostly invariant (with three exceptions in the word order database, all of which are in India and which allow numeral-noun order as well as noun-numeral order).

There are many additional issues raised which are beyond the scope of this study. One is how well Y DNA haplogroups would correlate with linguistic structures, or whether they would correlate with a different set of structures from mtDNA haplogroups. The finding of the paper is that genetic variation between populations predicts properties of unrelated languages, but in principle this can be genetic variation beyond that shown by mtDNA. MtDNA was chosen in particular partly because of the availability of the data across many populations, and partly in order to show that mtDNA can be informative about language history, *contra* the potential implication of Forster and Renfrew’s paper *(1)* that only Y DNA is informative about the spread of languages. Further research is needed to compare Y DNA haplogroups with structures.

An additional issue for which there is not sufficient space to explore is the history of how structures such as consonant cluster complexity may have spread between languages. Some likely events of language contact include Indo-European interaction with Uralic, Basque and Dravidian, as can be seen in the prevalence of Indo-European loan-words in these languages, while other language families such as those in the Caucasus may allow complex consonant clusters due to ancient relatedness with Indo-European. The likelihood of these different scenarios is another subject for further research, and in particular how the suggested spread of complex consonant clusters into Berber languages may have happened. A possible scenario is that Proto-Afro-Asiatic had complex codas, which were then subsequently lost in most branches but preserved in Berber. How plausible this is depends partly on the position of Berber in the Afro-Asiatic family, as it may have been a branch that split off early, causing the loss of complex consonant clusters to be an innovation common to the remaining branches of Afro-Asiatic. Alternatively, complex consonant clusters may have been brought into Berber languages by influence of Eurasian languages which are no longer spoken there. An additional possibility is that Berber languages innovated consonant clusters independently of Eurasian languages; it was suggested that this is unlikely given the fact that it correlates with haplogroup HV frequency in African languages taken by themselves, although what this might mean is that Berber languages have been sufficiently isolated from the influence of other African languages to innovate a different syllable structure from them, as well as to preserve high population frequencies of Eurasian haplogroups. The likelihood of these different scenarios can be evaluated by using a larger sample of linguistic structures and investigating the way that they have changed in the history of the Afro-Asiatic family.

Supplementary Figures S1-S8:

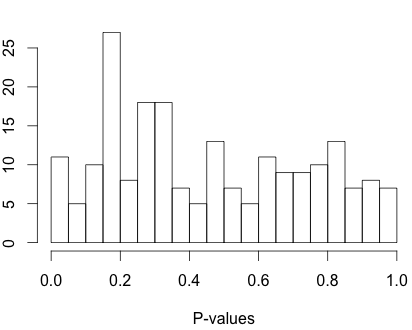


Figure S1: histogram of p-values of a random arrangement of the coda complexity data against all 252 haplogroups. Some tests have p-values below 0.05, but no more than have p-values in other intervals (e.g. the most is 0.15<p<0.2), suggesting that significant correlations in this data are purely due to multiple testing.

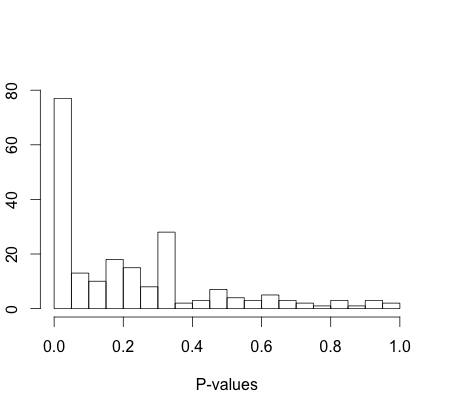
****

Figure S2: the histogram generated from comparing coda complexity with all 252 haplogroups. In this data, there are genuine correlations between coda complexity and some haplogroups, as shown by the fact that there are many more correlations with p-values below 0.05 than with p-values in other intervals.

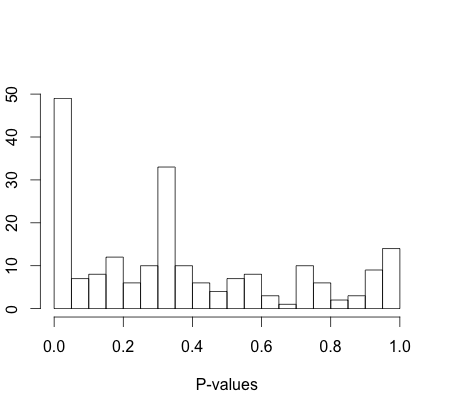
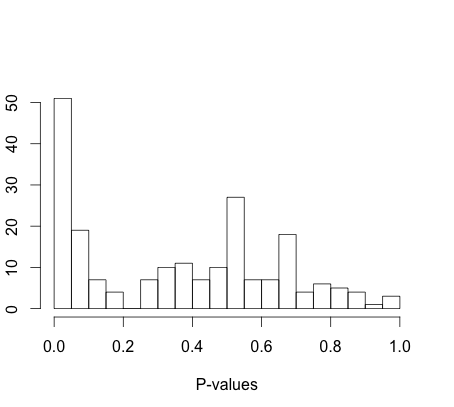
 

Figure S3: comparing mtDNA haplogroups with onset complexity (left) and number of tones (right).

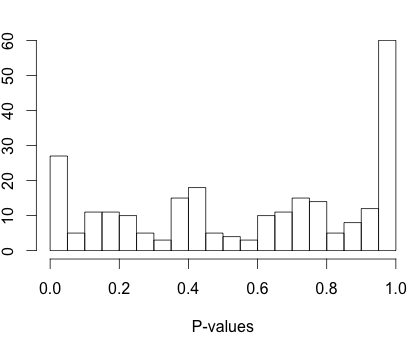


Figure S4: the p-values for mixed effects models for mtDNA haplogroups and coda complexity.

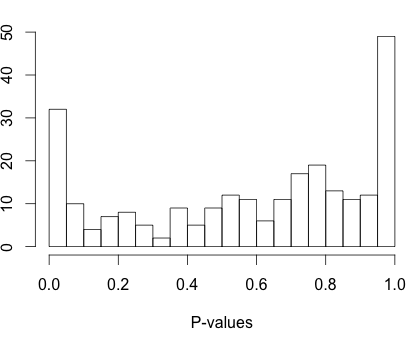
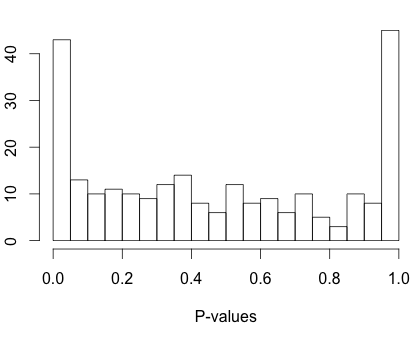


Figure S5: mixed effects models for onset complexity (left) and number of tones (right).

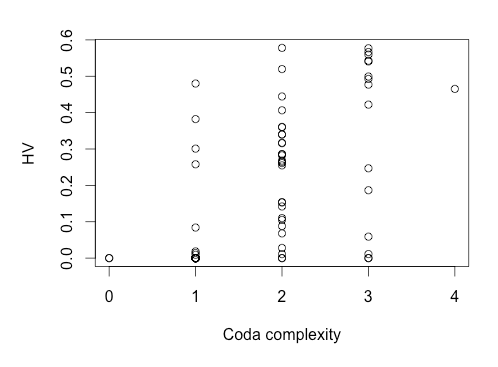


Figure S6: plot of coda complexity (number of consonants allowed in the coda of syllables) against haplogroup HV frequency.

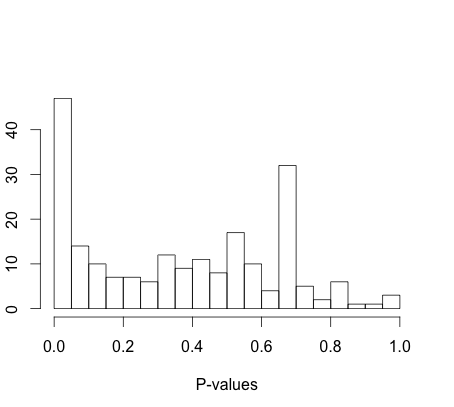
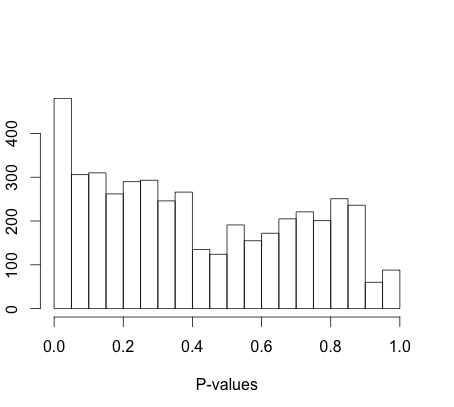


Figure S7: comparing mtDNA haplogroups with all word orders (left) and numeral-noun order in particular (right).

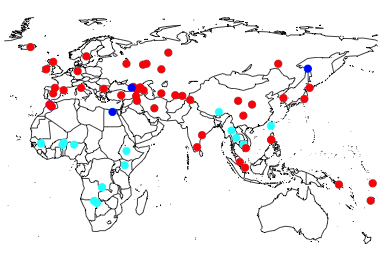


Fig. S8: The spread of numeral-noun word order: languages in red place the numeral before the noun; languages in dark blue have the option of placing the numeral before the noun; and languages in light blue place the numeral after the noun.