Word Structure Differences Reflect Genetic Distances Between Populations in Unrelated Language Families

Word Structures Reflect the History of Populations Speaking Unrelated Languages Spreading in Africa and Eurasia

There are around 30 different language families in Africa and Eurasia, with no consensus on how to relate them. Beyond vocabulary, one can use the structure of words. Rules such as whether a language allows consonant clusters can be stable over time because it is a general property of the lexicon. Syllable structure properties are expected to correlate with mitochondrial DNA through inheritance and substrate effects. Here it is shown that this holds when controlling for known language families and geographic distance. These spreads correspond to known agricultural expansions.

However, language families can share similar structural properties if they are in the same region; many languages in Southeast Asia are tonal, for example, despite being unrelated. This paper shows that syllable structures correlate with mitochondrial DNA haplogroups across multiple language families of Africa and Eurasia, after controlling for known language relatedness and geographic distance, suggesting that events such as the spread of agriculture also spread languages with particular syllable structures through different regions.

Languages vary in the structures of words that they allow. Georgian allows words with several consecutive consonants such as *mtsvane*, while Japanese which mostly only allows words with a single consonant in a sequence (apart from ‘n’) such as *hajimaru* ‘begin’. Languages in a similar region converge on similar syllable structures.

One possible way that this happens is by substrate influence. If a language arrives in a new region, then it is learnt by a population which was already present in the region. Chinese or Malay speakers in Singapore learn English and simplify some consonant clusters such as *lisp* to *lis* (*5*), because Chinese and Malay lack these consonant clusters. Conversely, some second-language learners can complexify the language that they are learning. Moroccan Arabic has developed complex consonant clusters where Standard Arabic does not have them, such as by changing Standard Arabic *kitabtu* to *ktebt*, and this is likely to be under the influence of the original local Berber languages which themselves allow very complex consonant clusters (*6*).

This can be shown by comparing syllable structures with mitochondrial DNA haplogroups.

Syllable structure boundaries closely match genetic boundaries. For example, Berber populations are genetically distinct from other populations within Africa, and have a distinctive syllable structure.

A test of this hypothesis is to compare linguistic structures with genetic variants whose movements can be tracked and dated, such as mitochondrial DNA haplogroups, and to see to what extent their distributions correlate. These correlation can then be retested while controlling for language family in order to see whether these hold after factoring out known relatedness of languages; and one can also control for geographic distance to see whether any similarity in distributions is just due to both genes and linguistic structures being spatially auto-correlated.

Syllable structures in particular give evidence for the spread of complex consonant cluster languages in Europe, India and North Africa, which correlate strongly with haplogroups such as HV after controlling for known relatedness and geographical distance, suggesting that they spread into these areas during the early Neolithic from Southwest Asia.

This paper shows that structural variation between languages of Africa and Eurasia predicts population spreads that correlate with mitochondrial DNA haplogroup frequencies, a correlation which remains after controlling for known language relatedness. Some structural properties have travelled across the whole of Eurasia through thirteen language families, and including into Berber languages of North Africa, giving linguistic evidence for population expansions that spread agriculture in Eurasia.

This hypothesis was tested on 73 populations in Africa and Eurasia up to the Pacific, from twenty-six different language families. Population frequencies of 252 haplogroups were collected from the genetics literature on these populations *(7-20)*.

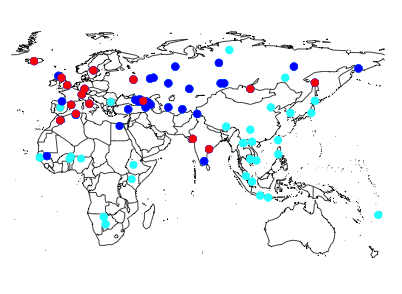
Structures were tested from three linguistic domains: syllable structures(Three syllable structure features were tested (how many consonants are allowed at the beginning of a syllable, how many at the end, and how many tones are used), using data from an independently coded database *(21)*). Mantel tests were used controlling for family and distance. Syllable structures correlated significantly; The first method was to use a Mantel test of differences in frequencies of haplogroups with differences in these three syllable structure features, the results of which were significant for coda complexity (r=0.35, p=0.001) and number of tones (r=0.2, p=0.03) but not onset complexity (p=0.17); using partial Mantel tests one can also simultaneously control for language family and geographic distance, for which the results are still significant (coda complexity r=0.14, p=0.003; number of tones r=0.18, p=0.008). Genetic distance in fact correlates with these two features better than language families do (r=0.14 for coda complexity and r=0.15 for number of tones).

The strongest correlation is between coda complexity and haplogroup HV (R2=0.34, p=3.8e-08), a correlation that is significant after a Bonferroni correction for the 756 tests done. Haplogroup HV is a large haplogroup found in Europe, North Africa, and decreasing in frequency in eastern Eurasia. This distribution is similar to the distribution of languages that allow complex consonant clusters in the coda of syllables (Figs. 1-2). In order to test whether this was more than expected due to known language families, a mixed effects model with haplogroup HV frequency and random intercepts for language family was used, which predicted coda complexity significantly better than a model with just random intercepts for language family (χ2=5.39, p=0.02). A second method was to randomly sample one language per known language family and retest the correlation between haplogroup HV and coda complexity, which was significant in 49% of the random samples. Haplogroup HV also correlates with other syllabic features such as the complexity of syllable onsets (R2=0.19, p=7.4e-05) and inversely with the number of tonal distinctions (R2=0.13, p=0.013), and both are similarly significant in mixed effects models (p<0.05).

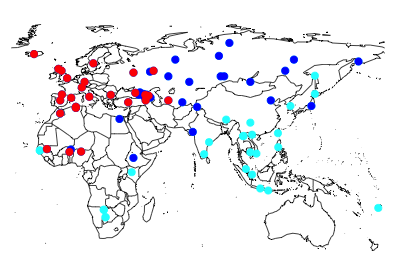
Complex consonant clusters have been traveling between thirteen language families in Eurasia such as Indo-European, Uralic, Northwest Caucasian, Basque, Turkic, Nakh-Daghestanian, Yeniseic and Dravidian along a similar path to that of the spread of haplogroup HV, either because these language families have been in contact or in some cases because they may be distantly related. The most complex consonant clusters in Eurasia are found in Indo-European and language families of the Caucasus, consistent with the high frequencies of haplogroup HV in European and Caucasian populations, suggesting that this property was innovated in Southwest Asia and spread out from there. Complex consonant clusters and haplogroup HV furthermore show a similar distribution within Africa. Haplogroup HV correlates significantly within African languages taken by themselves with coda complexity (R2=0.67, p=0.015) and inversely with the number of tonal distinctions (R2=0.78, p=0.006), although not with onset complexity (R2=0.34, p=0.1). On an expanded sample of 13 African languages for which haplogroup HV data was available *(22)*, all three of these correlations with haplogroup HV were significant (e.g. with codas, R2=0.82, p=2e-05). The highest frequencies of HV and the most complex syllables are found in North Africa in Berber populations of Morocco and Algeria. The syllable complexity of Berber languages is not likely to be because of Indo-European influence (such as from Latin or French), because Berber languages allow syllables more complex than those found in Indo-European languages (often even without any vowels such as *tkst* 'you feed on'), and these are primarily in indigenous words that are not known to be borrowed or influenced by any Indo-European language *(23)*; nor were these consonant clusters introduced by Arabic, which does not allow them. Instead, Berber languages seem to have a Eurasian affinity, in the same way that Berber populations have high frequencies of Eurasian mtDNA haplogroups (e.g. haplogroup U and haplogroup J, which also correlate with complex codas), lineages which came into North Africa from the Near East or Iberia during the early Neolithic *(24)*. The distinctiveness of Berber syllable structure and the fact that it correlates with Eurasian haplogroups suggests that it was brought by Neolithic Eurasian migration into North Africa, for which there is clear archeological and genetic evidence, but which has so far been lacking in linguistics *(25).*

The pattern that emerges out of these linguistic features is that structures have been spreading in Eurasia and North Africa, but not into the rest of Africa or Southeast Asia, resembling the expansion of Eurasian agriculture such as wheat and lactose tolerance. This result also shows the way that languages are non-independent even when they are in unrelated families, casting doubt on recent research that has proposed that tonal languages are adaptive in warm climates, or that tone is caused by low frequencies of particular genes controlling brain growth. The result of this paper suggests that the skewed geographical distributions of structures such as tone is an artifact of large-scale migration, rather than the causal influence of genes or warm climate.

Nature submission info: jeremycollinsmpi, bpasbpas



**Fig. 1**. The spread of complex syllables: languages in red allow three or more consonants in the coda of syllables; languages in dark blue allow two consonants; languages in light blue allow fewer than two consonants.



**Fig. 2.** The spread of haplogroup HV: populations in red have a population frequency of haplogroup HV above 30%; populations in dark blue have a frequency above 0%; and populations in light blue do not have haplogroup HV.