**Mitochondrial DNA Tracks The Neolithic Spread of Word Structures In Unrelated Languages Of Eurasia and Africa**

**Author:** Jeremy Collins1,2\*

**Affiliations:**

1Radboud University, Nijmegen, The Netherlands.

2International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands.

\*Correspondence to jeremy.collins@mpi.nl

Words in a language die out over time, but the patterns that words conform to can remain the same for much longer: for example, Indo-European languages often allow unusually complex consonant clusters, while Sino-Tibetan languages do not. The distribution of these word structures potentially indicates the way that languages have spread in prehistory, beyond what is known through language families. This is shown by the fact that word structures correlate with mitochondrial DNA variation after controlling for known language relatedness. The correlations between structures and particular haplogroups show the way that these features have spread geographically and when these structures are likely to have been innovated. Using this method, the spread of tone in Southeast Asia is dated to 7655.8 ± 3043.2 years ago in southern China by comparison with haplogroup F1a1a, and the spread of complex syllable codas is dated to 9739.5 ± 1357.6 years ago in southwest Asia by comparison with haplogroup V. This is a new method of investigating linguistic prehistory and comparing it with genetics and archeology, by dating and placing the origin of structural properties rather than language families.

There are over thirty language families in Africa and Eurasia, which cannot currently be shown to be related beyond approximately 8,000 years *(1)*. This paper tests the hypothesis that the structure of words may be informative about language history beyond this time-depth. The reason is that word-formation rules are a property of the whole lexicon: individual words die out, but new words are constantly being created by word formation rules (such as compounding: *foot-ball* or affixation: *global-isation*) that conserve the sound patterns that words conform to. Accordingly, many language families do have distinctive word structures: Uralic languages typically do not allow consonant clusters, while Indo-European languages do (such as *strength*); other families use tone to distinguish words such as Sino-Tibetan languages. The tendency for families to be homogenous makes these rules akin to some of the most stable vocabulary items, retained from an ancestral language with high probability despite thousands of years of evolutionary history in each branch of the family *(2)*.

The question is whether word structures can be informative about language history beyond what we know through the history of language families. In order to answer this question, word structures were compared with genetic variation in seventy-four populations in Africa and Eurasia up to the Pacific, from twenty-seven different language families. Mitochondrial DNA haplogroups were used because of the availability of this data for a large number of language families in Africa and Eurasia, using 252 haplogroups from the genetics literature *(3-16)*. Three word structure features were tested using data from an independently coded database, the World Phonotactics Database *(17):* how many consonants are allowed at the beginning (onset) of a syllable, how many at the end (coda), and how many tones are used. These three syllable features were tested in linear regressions with all 252 haplogroups. The way to correct for multiple testing was to take the p-values of all the linear regressions and find how many p-values were in between 0 and 0.05 compared with p-values at other intervals (e.g. between 0.4 and 0.45). If there are genuine correlations between word structures and haplogroups above that expected by multiple testing, then there should be many more below p-values below 0.05 than in other intervals. This pattern holds for the actual data, and does not hold for a random arrangement of the data (figures S1-S2). Mixed effects models were used with language family as a random intercept and comparing with a base model using just language family as a random effect, and for all three features these models were significantly better than the baseline model more times than expected by chance. The third method of correcting for multiple testing was to use a Mantel test of overall difference in haplogroup frequencies between populations and differences in syllable structures, 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 result is therefore that there are significant correlations between word structures and mtDNA haplogroups that cross-cut language families. The next question is whether specific population movements can be identified that caused word structures to spread; can we find when and where tonal languages originated, for example? Certain young haplogroups may have spread with tonal languages, and their coalescence times can place an upper limit on how long ago these expansions were. The coalescence times of haplogroups found in populations speaking tonal languages are summarized in table 1. Haplogroups B4c2 and F1a1a are both found across four language families of Southeast Asia, and are dated to 7109.9 ± 3676.6 years ago and 7655.8 ± 3043.2 years ago respectively (Figure 1). The mean dates agree with known dates of the expansion of Sino-Tibetan and Tai-Kadai from southern China, as well as the expansion of rice farming *(19)*.

Coda complexity is the other feature that correlates significantly with mitochondrial DNA variation. Haplogroup V is found across five language families with codas of two or more syllables, and in particular in families with three or more syllable (Figure 2). This haplogroup is dated to 9739.5 ± 1357.6 years ago, and is found in a strikingly similar distribution to complex codas: northern and eastern Europe, the Caucasus, and Berber populations in North Africa *(20-22)*. This haplogroup and others that correlate with coda complexity (HV, U, K, T) have been found in ancient European farmers and have been linked with demic spread of farming from the Near East *(23)*. Complex codas are hence likely to have been innovated in Southwest Asia in the early Neolithic and spread into Europe and North Africa with Indo-European and Afro-Asiatic languages. This account fits with the archeological evidence for the way that the ancestors of Indo-European and Afro-Asiatic speakers migrated. A linguistic ancestor of Indo-European languages is likely to have come from Anatolia *(19)*, if not Proto-Indo-European itself as has been argued in a recent study *(24)*. Similarly, there were Neolithic migrations into North Africa from southwest Asia as shown by the Afro-Asiatic speakers, and Berbers in particular who have high frequencies of Eurasian Y chromosome and mtDNA haplogroups, as well as high lactose tolerance; however, current linguistic comparison can only reconstruct the Afro-Asiatic language family to a likely origin in east Africa *(19)*. In the scenario outlined above, the use of complex consonant clusters in Berber languages is evidence for the ultimate origin of an ancestor of Afro-Asiatic languages in southwest Asia. To test the robustness of this claim, an expanded sample of African languages was taken for which data on haplogroup HV0 and haplogroup U data were available *(25)*, and these haplogroups were found to correlate significantly with coda complexity in African languages by themselves (e.g. with haplogroup HV0, R2=0.6, p=0.0002), showing that this linguistic property is associated with Eurasian influence in North Africa. It should be emphasized that this is not due to modern European or Arabic influence in North Africa, as Berber languages typically allow words with consonant clusters more complex than Arabic or any Indo-European language, often even without any vowels, such as *tkst* 'you feed on' *(26)*.

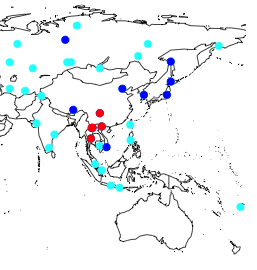
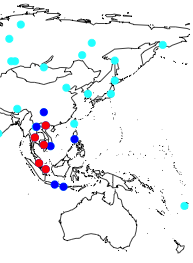
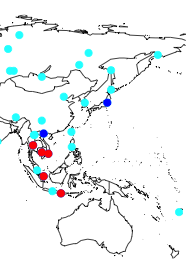
Two methods were used to show that the correlations discussed in this paper are not due to known language relatedness. The first method was to use a linear mixed effects model with the frequency of a particular haplogroup as a fixed effect and random intercepts for language family, and to see if it predicted the relevant word structure significantly better than a model with just random intercepts for language family. This held for the correlation between tone and haplogroups such as B5a (p<0.00001) and F2 (p<0.05) although narrowly not for F1a1a (p=0.06). The correlations between coda complexity and haplogroups HV0, V, T and U remain significant (e.g. haplogroup V, p<0.0001). The second method was to sample one language per family and to retest the correlations on one hundred random samples. The correlations hold up in many more random samples than that expected by chance: haplogroup V correlates with significantly with coda complexity in 38% of random samples, haplogroup T and haplogroup HV0 in 26% and 27%, and haplogroup U in 23%. Tone correlates with B5a in 94% of random samples, F2 in 70% and F1a1a in 29%.

There are two ways that these correlations arise. One is if language families are related at a greater time-depth than we can currently reconstruct, leading genes and linguistic properties to be inherited from a common ancestor. The other is if word structures are borrowed across languages when people migrate, for example in the way that Vietnamese was restructured by migration of Chinese people in Vietnam *(27)*. In both cases, the correlations with mtDNA haplogroups suggest that the word structures are carried by movement of people rather than by purely cultural transmission.

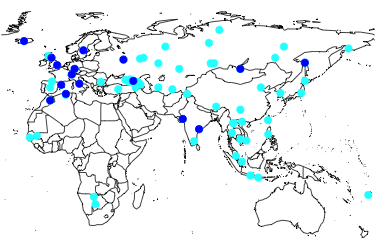
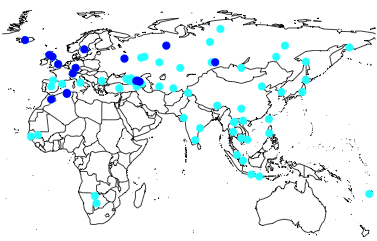
**References and Notes:**

1. H. Hammarström, R. Forkel, M. Haspelmath, S. Nordhoff, Glottolog 2.0. (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2014).
2. M. Pagel M, Q. D. Atkinson, A. Meade, Frequency of word use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717-720 (2007)
3. C. Barbieri *et al.*, Ancient substructure in early mtDNA lineages of Southern Africa. *Am. J. Hum. Genet.* **92(2)**, 285-292 (2013)
4. C. Barbieri *et al.*, Contrasting Maternal and Paternal Histories in the Linguistic Context of Burkina Faso. Mol. Biol. Evol.**29(4)**, 1213-1223. (2012)
5. M. Derenko *et al*, Phylogeographic analysis of mitochondrial DNA in Northern Asian populations. *Am. J. Hum. Genet.* **81(5)**, 1025-1041 (2007)
6. T. Duggan *et al.*, Maternal History of Oceania from Complete mtDNA Genomes: Contrasting Ancient Diversity with Recent Homogenization Due to the Austronesian Expansion. *Am. J. Hum. Genet.* **94(5)**, 721-733. (2014)
7. S. A. Federova *et al.*, Analysis of Mitochondrial DNA Lineages in Yakuts. *Molecular Biology* **37**, 544-553 (2003)
8. Helgason *et al.*, mtDNA and the Islands of the North Atlantic: Estimating the Proportions of Norse and Gaelic Ancestry, *Am. J. Hum. Genet.* **68**, 723-737 (2001)
9. T. Kivisild *et al.*, The Genetic Heritage of the Earliest Settlers Persists Both in Indian Tribal and Caste Populations. *Am. J. Hum. Genet.* **2(2)**, 313-332 (2003)
10. M. S. Peng *et al.*, Tracing the Austronesian Footprint in Mainland Southeast Asia: A Perspective from Mitochondrial DNA. *Mol. Biol. Evol.* **27(10)**, 2417-2430 (2010)
11. S. Plaza *et al.*, Joining the Pillars of Hercules, mtDNA Sequences Show Multidirectional Gene Flow in the Western Mediterranean, *Ann. Hum. Genet.* **67**, 312-328 (2003)
12. L. Quintana-Murci *et al.*,Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845 (2004)
13. E. B. Starikovskaya *et al.*, Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. *Ann. Hum. Genet.* **69**, 67-89 (2005)
14. M. Tanaka *et al.*,Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.* **14**, 1832–1850 (2004)
15. B. Wen *et al.*, Genetic Structure of Hmong-Mien Speaking Populations in East Asia as Revealed by mtDNA Lineages. *Mol. Biol. Evol.* **22(3)**, 725-734 (2005)
16. B. Yunusbayev *et al.*, The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* **29(1)**, 359-365 (2012)
17. M. Donohue, R. Hetherington, J. McElvenny, V. Dawson, World Phonotactics Database. 2013. Department of Linguistics, The Australian National University (2013) <http://phonotactics.anu.edu.au>. Accessed 1/10/2014.
18. D.M. Behar *et al.*, A “Copernican” Reassessment of the Human Mitochondrial DNA Tree from its Root. *Am. J. Hum. Genet.* **90(4)**, 675-84 (2012).
19. P. Bellwood, *First Migrants: Ancient Migration in Global Perspective.* (Wiley, 2013)
20. A.Bekada *et al.*, Introducing the Algerian Mitochondrial DNA and Y-Chromosome Profiles into the North African Landscape. *PLoS One* **8(2)**, e56775 (2013)
21. K. Fadhlaoui-Zid, *et al.*, Mitochondrial DNA heterogeneity in Tunisian Berbers. *Annals of Human Genetics* **68(3)**, 322–233 (2004).
22. C. Coudray, *et al.*, The Complex and Diversified Mitochondrial Gene Pool of Berber Populations. *Annals of Human Genetics* **73**, 196-214 (2009).
23. W. Haak *et al*., Ancient DNA from the First European Farmers in 7500-Year-Old Neolithic Sites. *Science* **310**, 1016-1018 (2005).
24. R. Bouckaert, *et al.*, Mapping the Origins and Expansion of the Indo-European Language Family. *Science* **337**, 957-960 (2012).
25. D. A. Badro *et al.*, Y-Chromosome and mtDNA Genetics Reveal Significant Contrasts in Affinities of Modern Middle Eastern Populations with European and African Populations. *PLoS ONE* **8(1)**, e54616. (2013)
26. R. Ridouane, Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber. *Phonology* **25**, 321-359 (2008)
27. N. J. Enfield, Areal linguistics and mainland Southeast Asia. *Annual Review of Anthropology* **34**, 181-206 (2005)
28. P. Forster, C. Renfrew, Mother tongue and Y chromosomes. *Science* **333,** 1390-1391 (2011)
29. N. Creanza, *et al.*, A comparison of worldwide phonemic and genetic variation in human populations. *PNAS* **112**, 1265-1272.
30. M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30(2)**, E386-E394. (2009) <http://www.phylotree.org>.

**Acknowledgements:** Thanks to C. Barbieri, D. Blasi, D. Dediu, M. Dunn, N. Enfield, S. Graham, R. Hunsucker, P. Muysken, E. Ritten, S. Roberts and K. Smiet for feedback. This work is supported by the Centre for Language Studies, Radboud University, Nijmegen.

**Fig. 1**. The spread of tonal languages in Southeast Asia (left), and the spread of haplogroups F1a1a (middle) and B4c2 (right). Languages in red allow more than three tones, languages in dark blue have tone, and languages in light blue lack tone. Populations in red have the highest frequencies of haplogroup F1a1a and B4c2, with decreasing frequencies in populations in dark blue, and absent in populations in light blue.

**Fig. 2**. The spread of languages with three or more consonants in the coda of syllables (left), and the spread of haplogroup V (right).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Number of language families | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Tone | 4333.9 | 5381.6 | 7047.2 | 7109.9 | 10863.5 | 10863.5 | 17520.4 |
| Complex codas | 1268.2 | 1268.2 | 9739.5 | 9739.5 | 9739.5 | 12846 | 12846 |

**Table 1.** Mean coalescence times of the youngest haplogroup that is found in tonal language families (row 1) and families with complex codas (row 2). The columns show the number of families that the haplogroup is found in.