

Relation between Word Order of Languages and the Entropy of Mitochondrial DNA Haplogroups Distribution of the Speakers' Population in Eurasia

ユーラシア大陸における言語の語順とその話者集団の ミトコンドリア DNA ハプログループ分布のエントロピーとの関係

Terumasa EHARA
江原暉将

Ehara NLP Research Laboratory
江原自然言語処理研究室
<http://www.ne.jp/asahi/eharate/eharate/>

1 Introduction

We are investigating the relation between the word order of languages (we focus on OV: object verb versus VO: verb object) and the speakers' thought pattern. We have approached it through diversity of DNA of the speakers' population.

Higher diversity of DNA means that the population has the cooperative thought pattern and lower diversity of DNA means that the population has the competitive thought pattern.

The diversity of DNA was measured by the entropy of haplogroup (Hg) distribution. Our conjecture is that OV language speakers' population has higher entropy value and VO language speakers' population has lower entropy value.

Ehara (2018) showed that OV language speakers' population have higher entropy value of Y-chromosome DNA haplogroup (YHg) than VO language speakers' population. Ehara (2019) showed that OV language speakers' population have higher entropy value of Mitochondrial DNA haplogroup (MtHg) than VO type language speakers' population.

Ehara (2019) treated the data only from Europe and the surrounding area. In this paper, we will analyze the data from larger area, Eurasia.

2 Data

Data for the word order features (OV/VO) are obtained from the WALS online database (Dryer, 2013) and many other web sites.

MtHg data of populations are obtained from 1000 Genomes Project (Rishishwar, 2017). From Rishishwar (2017), we can obtain the MtHg data of 2054 samples (individuals) from 26 worldwide populations. We estimate native languages spoken by these populations. We can recognize 22 languages in these data. We exclude the data that native languages aren't estimated. Populations and languages in our analysis are listed in Appendix 1 (Table A2).

The data include 26 macro MtHGs : $G = \{ L0, L1, L2, L3, L4, L5, D, F, G, N, Y, Z, H, I, J, K, T, V, W, X, U, M, R, A, B, C \}$ (see Table A1).

The entropy by Hgs for each language l : $H(l)$ is calculated by

$$H(l) = - \sum_{g \in G} p(l, g) \log_{n(G)} p(l, g)$$

where $p(l, g)$ is the probability (relative frequency of samples) of the Hg g for the language l and $n(G)$ is the number of Hgs (in our case $n(G)=26$).

3 Method of the analysis

We conduct t-test for the entropy value of OV and VO word order groups. Contrary to the conjecture, t-score of OV versus VO is negative (-1.76408) shown in Table 1 (a). It means that OV language speakers' population have lower diversity of DNA than VO language speakers' population (although this statement is not significant with significant level 5%). To deal with the problem, we will use the concept of "artificial

sub haplogroup”.

Table 1. T-test results

| | OV | VO |
|---------------|----------|--------|
| N | 9 | 13 |
| Mean | 0.4369 | 0.4985 |
| Unbiased var. | 0.0043 | 0.0059 |
| T | -1.96385 | |
| Two sided P | 0.0636 | |

(a) T-test result with macro HGs

| | OV | VO |
|---------------|---------|--------|
| N | 9 | 13 |
| Mean | 0.677 | 0.6507 |
| Unbiased var. | 0.125 | 0.0068 |
| T | 0.63732 | |
| Two sided P | 0.53114 | |

(b) T-test result with sub HGs

| | OV | VO |
|---------------|---------|--------|
| N | 6 | 8 |
| Mean | 0.7478 | 0.7071 |
| Unbiased var. | 0.001 | 0.0009 |
| T | 2.48515 | |
| Two sided P | 0.02869 | |

(c) T-test result with sub HGs only from Eurasia

| | OV | VO |
|---------------|---------|--------|
| N | 18 | 40 |
| Mean | 0.7917 | 0.7429 |
| Unbiased var. | 0.0029 | 0.0059 |
| T | 2.44314 | |
| Two sided P | 0.01774 | |

(d) T-test result with sub HGs from European and surrounding areas

3.1 Artificial sub haplogroup

Hgs listed in G are macro Hgs and they have different sizes. For example, Hg M is large and Y is small. We define the size of Hg by the number of mutations, insertions and deletions in the Hg. These numbers are obtained from PhyloTree¹ (van Oven and Kayser, 2008) and listed in Table 2.

Table 2. Macro haplogroups and their sizes

| Hg | A | B | C | D | E | F | G | H | HV |
|------|-----|------|------|-----|-----|-----|-----|------|-----|
| Size | 389 | 725 | 300 | 750 | 53 | 254 | 167 | 1395 | 181 |
| Hg | I | J | JT | K | L0 | L1 | L2 | L3 | L4 |
| Size | 131 | 443 | 3 | 355 | 615 | 421 | 368 | 564 | 99 |
| Hg | L5 | L6 | M | N | O | P | Q | R | S |
| Size | 108 | 27 | 2192 | 464 | 14 | 159 | 123 | 601 | 61 |
| Hg | T | U | V | W | X | Y | Z | | |
| Size | 393 | 1110 | 80 | 107 | 188 | 29 | 70 | | |

¹ Actual data are obtained from https://www.phylotree.org/builds/mtDNA_tree_Build_17.zip.

We define the number of sub Hgs (SHGs) of a macro Hg by the rounded up value of size/100. For example, the number of SHg of macro Hg M is 22 and macro Hg Y is 1. These SHGs are defined artificially. So, we call them “artificial sub haplogroups”.

3.2 Distribution within artificial sub haplogroups

We distribute relative frequency of the macro Hg g to its artificial SHGs. Uniform distribution may be too diverse. So, we use Gaussian like distribution within SHGs. Precisely, when $m(g)$ is the number of SHGs of a macro Hg g , i th SHG ($i=1, \dots, m(g)$) has the relative frequency $r(i, g)$:

$$r(i, g) = \begin{cases} cdf(-3 + 6i/m(g)) \times R(g) & (i = 1) \\ \{1 - cdf(-3 + 6(i - 1)/m(g))\} \times R(g) & (i = m) \\ \{cdf(-3 + 6i/m(g)) - cdf(-3 + 6(i - 1)/m(g))\} \times R(g) & (else) \end{cases}$$

where $cdf(\cdot)$ is the standard Gaussian cumulative distribution function and $R(g)$ is the relative frequency of macro Hg g .

The equation of entropy calculation is not changed except for G is the set of all artificial SHGs.

4 Results

4.1 Results from the data of the present paper

Using artificial sub haplogroups, we re-calculate the entropy for the data and re-conduct t-test. The result are shown in Table 1 (b). The T-value is positive (0.63732), however, two sided P is extremely large and the conjecture is not significant by these data.

We restrict the data only from Eurasia, excluding four data from Africa (Esan, Mandinca, Luhya and Yorba) and three data from America (Arawak, Zapotec/Chontal and Quechua/Aymara). The result is shown in Table 1 (c). The T-value is positive (2.48515) and two sided P is small enough which is significant with significant level 5%. The detail of the calculation results are shown in Appendix 1.

Figure 1 shows the ranking of the entropy values for each language speakers’ population from

lower to higher. African and American populations rather have lower entropy value.

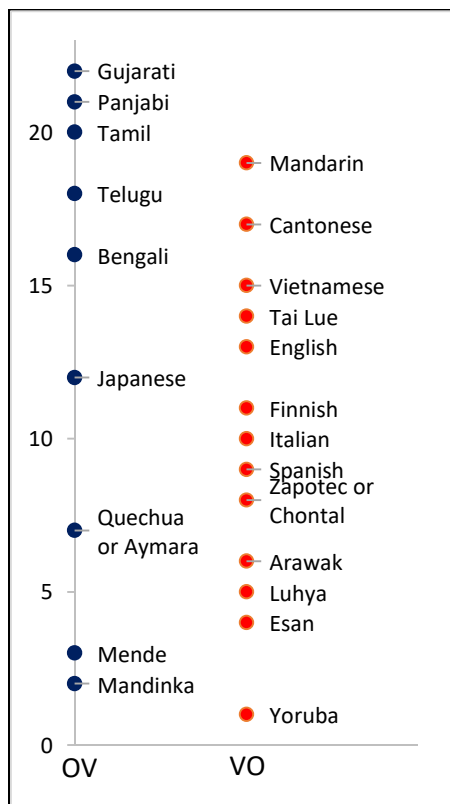


Figure 1. Ranking of artificial SHg entropy of each language speakers' population (lower to higher)

4.2 Results from the data of previous paper

Applying our artificial sub haplogrouping method to the data of previous paper (Ehara, 2019), we obtain the result shown in Table 1 (d). The T-value is positive (2.44314) and two sided P is small enough which is significant with significant level 5%.

5 Conclusion

Relation between word order (Object (O) / Verb (V)) of languages and the entropy of Mitochondrial DNA haplogroups distribution of the speakers' population is examined. T-test results using "artificial sub haplogroup" show OV word order language speakers' population tend to have higher entropy value than VO word order language speakers' population.

The languages used in this analysis are not exhaustive. The number of languages are only 14 from Eurasia. The number of languages in the world are more than several thousand.

Analysis using the data from these language speakers is the remaining issue.

References

- Matthew S. Dryer. 2013. Order of Object and Verb and Order of Adjective and Noun, *In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (<http://wals.info/chapter/83> and 87, Accessed on 2015-3-23).
- Terumasa Ehara. 2018. Relation between Word Order of Languages and the Entropy of Y-chromosome Haplogroup Distribution of the Speakers' Population, *Proceedings of The 24th Annual Meeting of The Association for Natural Language Processing*, P10-8, pp.1019-1022, March 2018.
- Terumasa Ehara. 2019. Relation between Word Order of Languages and the Entropy of Mitochondrial Haplogroups Distribution of the Speakers' Population, *Proceedings of The 25th Annual Meeting of The Association for Natural Language Processing*, P2-9, pp.490-493, March 2019.
- Lavanya Rishishwar and I. King Jordan. 2017. Implications of human evolution and admixture for mitochondrial replacement therapy, *BMC Genomics*, Vol. 18, pp.140-150, 2017.
- Mannis van Oven and Manfred Kayser. 2008. Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation, *HUMAN MUTATION Mutation in Brief*, #1039, 30:E386-E394, Sept. 2008.

Appendix 1 Base data used in the research

Base data used in the research are presented in the following 2 tables.

In Table A1, N means a number of samples. L0, L1, ... , C mean name of macro haplogroup (Hg) (cell values are relative frequency (%) of each Hg in the sample). The last row of Table A1 indicates the number of artificial sub haplogroups for each macro haplogroup.

In Table A2, feature values for O-V are OV, VO and blank (No data). Feature values for A-N (adjective versus noun) are AN, NA, NDO (No dominant order) and blank (No data). "E by Hg" means the entropy calculated by macro haplogroups and "E by SHg" means the entropy calculated by artificial sub haplogroups.

Table A1. Populations, macro haplogroups and relative frequencies (%)

| Area | Population | N | L0 | L1 | L2 | L3 | L4 | L5 | D | F | G | N | Y | Z | H | I | J | K | T | V | W | X | U | M | R | A | B | C | | | | |
|-----------|---|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Africa | Esan in Nigeria | 99 | 7.071 | 20.2 | 27.27 | 43.43 | 2.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | |
| | Gambian in Western Division, The Gambia | 113 | 0 | 13.27 | 42.48 | 37.17 | 1.77 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5.31 | 0 | 0 | 0 | 0 | 0 | | | |
| | Luhya in Webuye, Kenya | 101 | 17.82 | 7.921 | 11.88 | 46.54 | 4.95 | 9.901 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.99 | 0 | 0 | 0 | 0 | | |
| | Mende in Sierra Leone | 85 | 2.353 | 20 | 45.88 | 28.24 | 1.176 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.353 | 0 | 0 | 0 | 0 | 0 | | |
| | Yoruba in Ibadan, Nigeria | 108 | 4.63 | 15.74 | 35.19 | 43.52 | 0.926 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | |
| East Asia | Chinese Dai in Xishuangbanna, China | 99 | 0 | 0 | 0 | 0 | 0 | 0 | 12.12 | 25.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24.24 | 14.14 | 0 | 19.19 | 5.051 | | | |
| | Han Chinese in Beijing, China | 103 | 0 | 0 | 0 | 0 | 0 | 0 | 22.33 | 15.53 | 4.854 | 7.767 | 0 | 0.971 | 0 | 0 | 0 | 0 | 0.971 | 0 | 0 | 0 | 0 | 0 | 0 | 18.45 | 5.825 | 6.796 | 11.65 | 4.854 | | |
| | Southern Han Chinese, China | 108 | 0 | 0 | 0 | 0 | 0 | 0 | 21.3 | 14.82 | 1.852 | 9.259 | 0 | 2.778 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17.59 | 10.19 | 5.556 | 14.82 | 1.852 | | | |
| | Japanese in Tokyo, Japan | 104 | 0 | 0 | 0 | 0 | 0 | 0 | 37.5 | 5.769 | 10.58 | 8.654 | 0.962 | 3.846 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13.46 | 0 | 5.769 | 13.46 | 0 | | |
| | Kinh in Ho Chi Minh City, Vietnam | 101 | 0 | 0 | 0 | 0 | 0 | 0 | 1.98 | 26.73 | 0 | 3.96 | 0.99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 32.67 | 10.89 | 0.99 | 20.79 | 0.99 | | |
| Europe | Utah residents with NW European ancestry | 99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 51.52 | 1.01 | 8.081 | 3.03 | 10.1 | 3.03 | 5.051 | 0 | 18.18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | |
| | Finnish in Finland | 99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 37.37 | 2.02 | 7.071 | 6.061 | 3.03 | 4.04 | 2.02 | 2.02 | 36.36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | British in England and Scotland | 92 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 42.39 | 3.261 | 10.87 | 5.435 | 10.87 | 0 | 2.174 | 5.435 | 19.57 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | Iberian populations in Spain | 107 | 0 | 0.935 | 0 | 0.935 | 0 | 0 | 0 | 0 | 1.869 | 0 | 0 | 0 | 54.21 | 0.935 | 3.738 | 6.542 | 7.477 | 5.607 | 0 | 0 | 0 | 16.82 | 0 | 0.935 | 0 | 0 | 0 | 0 | 0 | |
| | Toscans in Italy | 108 | 0 | 0.926 | 0 | 0 | 0 | 0 | 0.926 | 0 | 0 | 0 | 0 | 50 | 0 | 7.407 | 8.333 | 12.04 | 1.852 | 3.704 | 0.926 | 13.89 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| India | Bengali in Bangladesh | 86 | 0 | 0 | 0 | 0 | 0 | 0 | 2.326 | 1.163 | 0 | 0 | 0 | 0 | 1.163 | 0 | 2.326 | 0 | 0 | 0 | 2.326 | 0 | 12.79 | 67.44 | 9.302 | 1.163 | 0 | 0 | 0 | 0 | 0 | |
| | Gujarati Indian in Houston, TX | 106 | 0 | 0 | 0.943 | 1.887 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6.604 | 0 | 0.943 | 0.943 | 1.887 | 0 | 1.887 | 1.887 | 14.15 | 38.68 | 30.19 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | Indian Telugu in the UK | 103 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.971 | 0 | 0 | 1.942 | 0.971 | 0.971 | 0.971 | 3.883 | 0 | 4.854 | 0 | 13.59 | 59.22 | 12.62 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Punjabi in Lahore, Pakistan | 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3.125 | 2.083 | 0 | 0 | 7.292 | 0 | 1.042 | 0 | 3.125 | 0 | 2.083 | 1.042 | 11.46 | 57.29 | 11.46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Sri Lankan Tamil in the UK | 103 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.913 | 0 | 0 | 11.65 | 0 | 0.971 | 0 | 0 | 0 | 0 | 0 | 0 | 13.59 | 48.54 | 21.36 | 0.971 | 0 | 0 | 0 | 0 | 0 | |
| America | African Caribbean in Barbados | 96 | 4.167 | 21.88 | 39.58 | 27.08 | 1.042 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.042 | 0 | 1.042 | 0 | 0 | 0 | 0 | 0 | 0 | 1.042 | 0 | 0 | 1.042 | 1.042 | 1.042 | 1.042 | 1.042 | 1.042 |
| | African Ancestry in Southwest US | 66 | 10.61 | 24.24 | 21.21 | 36.36 | 0 | 0 | 1.515 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.515 | 1.515 | 0 | 1.515 | 0 | 1.515 | 0 | 1.515 | |
| | Colombian in Medellin, Colombia | 94 | 1.064 | 2.128 | 2.128 | 4.255 | 0 | 0 | 2.128 | 0 | 0 | 0 | 0 | 0 | 1.064 | 0 | 0 | 0 | 2.128 | 0 | 0 | 0 | 0 | 0 | 1.064 | 0 | 42.55 | 35.11 | 6.383 | 0 | 0 | 0 |
| | Mexican Ancestry in Los Angeles, California | 67 | 0 | 0 | 0 | 0 | 0 | 0 | 13.43 | 0 | 0 | 0 | 0 | 0 | 7.463 | 0 | 0 | 0 | 0 | 0 | 1.493 | 1.493 | 0 | 2.985 | 0 | 0 | 37.31 | 22.39 | 13.43 | 0 | 0 | 0 |
| | Peruvian in Lima, Peru | 86 | 0 | 0 | 1.163 | 2.326 | 0 | 0 | 15.12 | 0 | 0 | 0 | 0 | 0 | 1.163 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16.28 | 46.51 | 17.44 | 0 | 0 | 0 | |
| | Puerto Rican in Puerto Rico | 105 | 1.905 | 4.762 | 3.81 | 9.524 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.905 | 0 | 5.714 | 0.952 | 0.952 | 0 | 0 | 0 | 0 | 3.81 | 0 | 0 | 36.19 | 6.667 | 23.81 | 0 | 0 | 0 | 0 |
| | Number of artificial SHg | --- | 7 | 5 | 4 | 6 | 1 | 2 | 8 | 3 | 2 | 6 | 1 | 1 | 14 | 2 | 5 | 4 | 4 | 1 | 2 | 2 | 12 | 24 | 10 | 4 | 8 | 4 | 4 | 4 | 4 | |

Table A2. Languages, word order and entropy

| Area | Population | Language | OV-VO | AN-NA | E by Hg | E by SHg |
|-----------|---|-----------------|-------|-------|---------|----------|
| Africa | Esan in Nigeria | Esan | VO | NDO | 0.401 | 0.525 |
| | Gambian in Western Division, The Gambia | Mandinka | OV | NA | 0.377 | 0.506 |
| | Luhya in Webuye, Kenya | Luhya | VO | NA | 0.473 | 0.570 |
| | Mende in Sierra Leone | Mende | OV | NA | 0.388 | 0.508 |
| | Yoruba in Ibadan, Nigeria | Yoruba | VO | NA | 0.370 | 0.503 |
| East Asia | Chinese Dai in Xishuangbanna, China | Tai Lue | VO | NA | 0.519 | 0.696 |
| | Han Chinese in Beijing, China | Mandarin | VO | AN | 0.650 | 0.756 |
| | Southern Han Chinese, China | Cantonese | VO | AN | 0.633 | 0.752 |
| | Japanese in Tokyo, Japan | Japanese | OV | AN | 0.570 | 0.690 |
| | Kinh in Ho Chi Minh City, Vietnam | Vietnamese | VO | NA | 0.500 | 0.696 |
| Europe | Utah residents with NW European ancestry | | | | 0.459 | 0.673 |
| | Finnish in Finland | Finnish | VO | AN | 0.480 | 0.689 |
| | British in England and Scotland | English | VO | AN | 0.515 | 0.695 |
| | Iberian populations in Spain | Spanish | VO | NA | 0.472 | 0.686 |
| | Toscans in Italy | Italian | VO | NA | 0.492 | 0.687 |
| India | Bengali in Bangladesh | Bengali | OV | AN | 0.358 | 0.735 |
| | Gujarati Indian in Houston, TX | Gujarati | OV | AN | 0.496 | 0.775 |
| | Indian Telugu in the UK | Telugu | OV | AN | 0.421 | 0.752 |
| | Punjabi in Lahore, Pakistan | Punjabi | OV | AN | 0.454 | 0.771 |
| | Sri Lankan Tamil in the UK | Tamil | OV | AN | 0.428 | 0.764 |
| America | African Caribbean in Barbados | | | | 0.466 | 0.562 |
| | African Ancestry in Southwest US | | | | 0.490 | 0.598 |
| | Colombian in Medellin, Colombia | Arawak | VO | NA | 0.465 | 0.583 |
| | Mexican Ancestry in Los Angeles, California | Zapotec/Chontal | VO | NA/AN | 0.511 | 0.621 |
| | Peruvian in Lima, Peru | Quechua/Aymara | OV | AN | 0.440 | 0.592 |
| | Puerto Rican in Puerto Rico | | | | 0.587 | 0.639 |