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

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 and the speakers' thought pattern. We have approached it through suicide rate and homicide rate (Ehara, 2015, 2017) and diversity of Y-chromosome haplogroups of the speakers' population (Ehara, 2018).

Y-chromosome DNA reflects paternal lineage and we measure its diversity with the entropy of Y-chromosome haplogroups (YHg) distribution. Higher entropy value means the population is rich in diversity and it relates to peaceful thought pattern. Lower entropy value means the population is poor in diversity and it relates to warlike thought pattern (Sakitani, 2008).

In this paper we use another metrics. It is the entropy of mitochondrial DNA haplogroups (MtHg) distribution of the speakers' population (shortly "MtHg entropy"). Mitochondrial DNA reflects maternal lineage.

Our conjecture is that head final language speakers' population has higher entropy value and head initial language speakers' population has lower entropy value. Ehara (2018) showed the conjecture is right for the YHg entropy.

2 Data

Data for the word order features are obtained from the WALS online database (Dryer, 2013). We use two dominant word order features:

- Order of Object (O) and Verb (V),
- Order of Adjective (A) and Noun (N).

WALS online provides O and V feature values for 1519 languages and A and N feature values for 1366 languages. Head initial languages tend to have VO and NA word order and head final languages tend to have OV and AN word order.

MtHg data of populations are obtained from Eupedia's page of "Distribution of European mitochondrial DNA (mtDNA) haplogroups by region in percentage" (Eupedia.com, 2018). From this page, we get the MtHg data from 80 populations in Europe and its surrounding regions. We estimate languages spoken by these populations. We can recognize 62 languages in these data. We merge population level data to language level data using the sample size of the population and each MtHg's relative frequency of the population. More concretely, the merging method is as follows. We use the set of all MtHg : $G = \{L, HV, H, HV0+V, J, T1, T2, U2, U3, U4, U5, U, K, I, W, X, Other\}$. Let P be the set of populations, L be the set of languages and $\mathcal{L} : P \rightarrow L$ be a mapping which corresponds each population to its speaking language. For each $l \in L$ and $g \in G$, the probability (relative frequency) of g for l : $p(l,g)$ is calculated by

$$p(l,g) = \frac{\sum_{p \in \mathcal{L}^{-1}(l)} q(p,g) \times s(p)}{\sum_{p \in \mathcal{L}^{-1}(l)} s(p)}$$

where $q(p,g)$ is the probability (relative frequency) of g for the population p and $s(p)$ is the sample size of the population p.

The MtHg entropy for each language l : $H(l)$

is calculated by

$$H(l) = -\sum_{g \in G} p(l,g) \log_{10} p(l,g)$$

Appendix 1 shows the base data used in the research sorted by the MtHg entropy.

3 Analysis and results

We conduct t-test for O and V word order groups and A and N word order groups. We discard "no dominant order" data. Results are shown in Table 1. For O and V case, the mean value of the entropy of the OV language group is significantly ($p<1\%$) higher than the mean value of the entropy of the VO language group. On the other hand, for A and N case, mean values are not different significantly ($p>1\%$). This results shows our conjecture is true by half for MtHg case. For YHg case, Ehara (2018) showed for both O and V case and A and N case, the mean values of the entropy values are significantly different. The mean value of the YHg entropy of OV language group is higher than the mean value of the YHg entropy of VO language group. The mean value of the YHg entropy of AN language group is higher than the mean value of the YHg entropy of NA language group.

Table 1: Results of t-test for the MtHg entropy

	OV	VO
n	17	36
mean	0.980	0.897
unbiased var.	0.007	0.012
t	2.748	
p	0.00828	

	AN	NA
n	38	19
mean	0.913	0.945
unbiased var.	0.011	0.013
t	-1.070	
p	0.2892	

Figure 1 shows the ranking of MtHg entropy values for each language from lower to higher. Many OV type languages have higher entropy values and many VO type languages have lower entropy values.

4 Correlation between MtHg entropy and YHg entropy

In this section, we compare the YHg entropy and the MtHg entropy. For 42 languages in the Appendix 1, YHg entropy can be obtained from Ehara (2018). They are listed in Appendix 2 with MtHg entropy and YHg entropy.

Scattering graph of MtHg entropy and YHg entropy is shown in Figure 2.

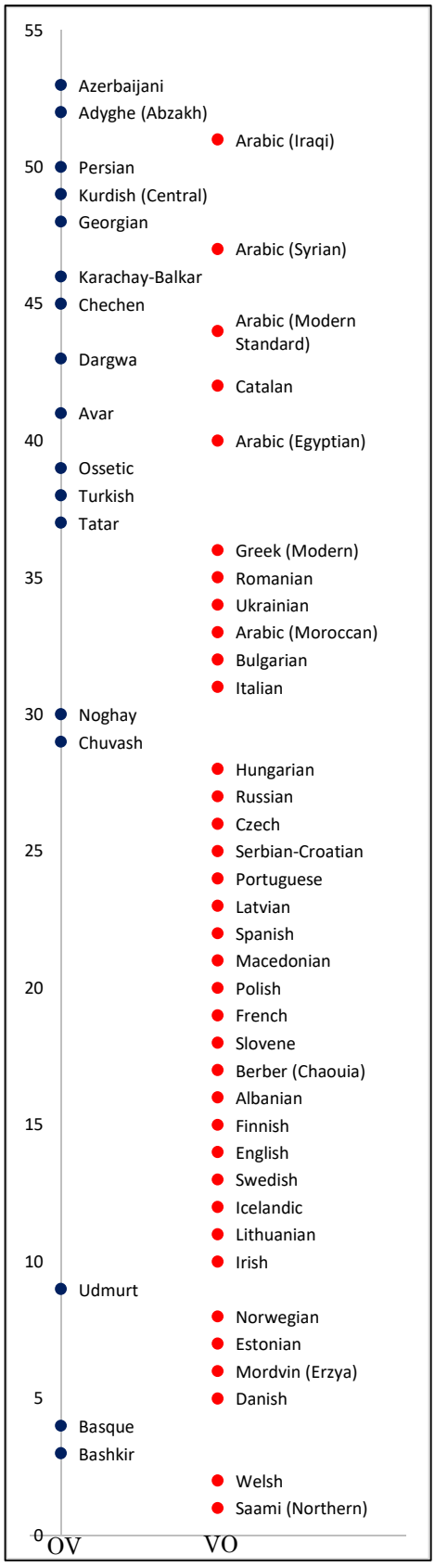

Figure 1: Ranking of MtHg entropy of each language (lower to higher)

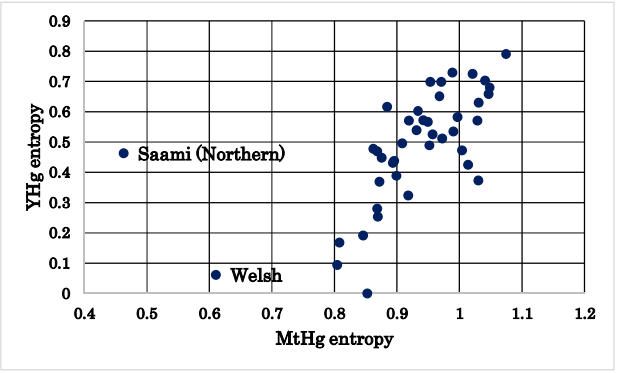

Figure 2: Scattering graph of MtHg entropy and YHg entorpy for 42 languages

Correlation coefficient of MtHg entropy and YHg entropy is 0.597. Excluding the outliers (Saami (Northern) and Welsh), correlation coefficient is 0.742.

T-test results for MtHg entropy and YHg entropy by restricted 42 languages are shown in Table 2. MtHg entropy values divided by OV and VO are significantly different (p<5%). However, MtHg entropy values divided by AN and NA, YHg entropy values divided by OV and VO and YHg entropy values divided by AN and NA are not significantly different.

Table 2: Results of t-test for the MtHg entropy and the YHg entropy

(a) MtHg entropy

	OV	VO
n	14	24
mean	0.973	0.893
standard dev.	0.091	0.122
t	2.154	
p	0.03805	

	AN	NA
n	29	13
mean	0.922	0.924
standard dev.	0.111	0.121
t	-0.038	
p	0.96985	

(b) YHg entropy

	OV	VO
n	14	24
mean	0.508	0.460
standard dev.	0.223	0.179
t	0.736	
p	0.46621	

	AN	NA
n	29	13
mean	0.495	0.455
standard dev.	0.169	0.232
t	0.631	
p	0.53179	

5 Conclusion

Relation between word order (Object (O) / Verb (V) and Adjective (A) / Noun (N)) of languages and the entropy of Mitochondrial DNA haplogroups distribution of the speakers' population is examined. T-test results 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. They are spoken in Europe and its surrounding regions. Analysis using languages all over the world is one of the remaining issues.

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. 2015. Relation between Word Order Parameters and Suicide / Homicide Rates, *Journal of Yamanashi Eiwa College*, Vol.13, pages 9-29.

Terumasa Ehara. 2017. Relation between the Word Order Characteristics and Suicide/Homicide Rates (6), *Proceedings of The 23th Annual Meeting of The Association for Natural Language Processing*, P4-4, pp.190-193.

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.

Eupedia.com. 2018. Distribution of European mitochondrial DNA (mtDNA) haplogroups by region in percentage, *Eupedia*. (https://www.eupedia.com/europe/european_mtdna_haplogroups_frequency.shtml, Accessed on 2018-12-5).

Mitsuru Sakitani. 2008. A long journey of Japanese populations revealed by DNA analysis, *Showado, Kyoto*, pages 37-39 (in Japanese).
崎谷満. 2008. DNAでたどる日本人10万年の旅, *昭和堂, 京都*, pages 37-39.

Appendix 1 Base data used in the research

Feature values for O and V are OV, VO, NDO (No dominant order), and blank (No data).
Feature values for A and N are AN, NA, and blank (No data).
Size means a number of the sample.
L, HV, ... , Other mean Mitochondrial DNA haplogroup (MtHg) names (cell values are relative frequency of MtHg in the sample).

Language	L	HV	H	HV0+V	J	T1	T2	U2	U3	U4	U5	U	K	I	W	X	Other	size	entropy	O-V	A-N
Saami (Northern)	0.000	0.000	0.038	0.417	0.000	0.000	0.004	0.000	0.000	0.000	0.481	0.000	0.000	0.000	0.014	0.000	0.046	499	0.4625	VO	AN
Welsh	0.000	0.000	0.598	0.044	0.152	0.022	0.011	0.000	0.000	0.000	0.044	0.000	0.076	0.033	0.000	0.011	0.011	92	0.6102	VO	NA
Galician	0.038	0.011	0.595	0.038	0.087	0.011	0.038	0.016	0.000	0.005	0.054	0.016	0.049	0.005	0.016	0.011	0.005	185	0.7076		
Bashkir	0.000	0.000	0.251	0.024	0.068	0.003	0.023	0.026	0.000	0.168	0.084	0.000	0.027	0.006	0.008	0.000	0.317	666	0.8044	OV	AN
Basque	0.003	0.008	0.490	0.079	0.076	0.015	0.060	0.010	0.003	0.008	0.117	0.019	0.053	0.007	0.011	0.023	0.018	618	0.8082	OV	NA
Danish	0.005	0.000	0.473	0.036	0.134	0.013	0.058	0.027	0.000	0.022	0.058	0.013	0.089	0.022	0.009	0.009	0.031	224	0.8197	VO	AN
Mordvin (Erzya)	0.000	0.010	0.422	0.049	0.078	0.020	0.078	0.069	0.000	0.020	0.157	0.020	0.000	0.059	0.000	0.000	0.020	102	0.8278	VO	AN
Estonian	0.000	0.008	0.458	0.015	0.107	0.015	0.076	0.015	0.008	0.057	0.134	0.012	0.019	0.012	0.031	0.008	0.027	262	0.8295	VO	AN
Norwegian	0.003	0.002	0.457	0.038	0.105	0.010	0.076	0.002	0.013	0.027	0.115	0.019	0.054	0.019	0.018	0.005	0.038	628	0.8386	VO	AN
Udmurt	0.000	0.000	0.218	0.000	0.020	0.149	0.238	0.099	0.000	0.040	0.089	0.020	0.000	0.000	0.000	0.000	0.129	101	0.8461	OV	AN
Irish	0.000	0.013	0.442	0.057	0.107	0.013	0.054	0.013	0.010	0.013	0.084	0.003	0.120	0.030	0.023	0.007	0.013	299	0.8525	VO	NA
Lithuanian	0.000	0.009	0.453	0.050	0.064	0.032	0.070	0.000	0.012	0.038	0.114	0.012	0.015	0.032	0.021	0.009	0.070	342	0.8594	VO	AN
Dutch	0.003	0.000	0.409	0.073	0.100	0.018	0.109	0.014	0.000	0.059	0.068	0.005	0.091	0.023	0.023	0.005	0.002	1000	0.8620	NDO	AN
Icelandic	0.002	0.037	0.386	0.021	0.140	0.005	0.103	0.000	0.000	0.029	0.079	0.002	0.100	0.040	0.005	0.015	0.034	998	0.8686	VO	AN
Swedish	0.006	0.005	0.458	0.050	0.077	0.027	0.041	0.008	0.006	0.030	0.121	0.028	0.064	0.028	0.013	0.013	0.025	637	0.8686	VO	AN
English	0.001	0.001	0.444	0.031	0.120	0.019	0.061	0.014	0.008	0.025	0.087	0.026	0.074	0.040	0.009	0.020	0.021	4186	0.8690	VO	AN
Finnish	0.000	0.000	0.363	0.073	0.059	0.020	0.024	0.006	0.000	0.011	0.207	0.008	0.045	0.042	0.096	0.013	0.033	971	0.8719	VO	AN
German	0.003	0.005	0.451	0.038	0.092	0.029	0.079	0.011	0.010	0.031	0.086	0.008	0.067	0.019	0.017	0.012	0.041	3089	0.8753	NDO	AN
Albanian	0.000	0.007	0.477	0.034	0.074	0.040	0.054	0.020	0.007	0.027	0.067	0.027	0.054	0.020	0.027	0.020	0.054	149	0.8841	VO	NA
Bosnian	0.006	0.012	0.459	0.048	0.083	0.017	0.044	0.008	0.008	0.050	0.088	0.012	0.050	0.033	0.029	0.008	0.050	521	0.8853		
Berber (Chaouia)	0.250	0.013	0.316	0.076	0.045	0.037	0.015	0.006	0.003	0.003	0.021	0.066	0.034	0.001	0.003	0.013	0.098	939	0.8862	VO	NA
Yiddish	0.014	0.050	0.229	0.041	0.068	0.000	0.048	0.000	0.010	0.010	0.020	0.022	0.318	0.013	0.016	0.008	0.134	836	0.8918		
Slovene	0.000	0.011	0.444	0.048	0.098	0.025	0.064	0.011	0.017	0.028	0.089	0.014	0.053	0.028	0.022	0.011	0.036	358	0.8934	VO	AN
French	0.007	0.015	0.449	0.048	0.074	0.019	0.069	0.014	0.013	0.024	0.082	0.018	0.085	0.022	0.023	0.009	0.030	1663	0.8955	VO	NA
Slovak	0.003	0.019	0.427	0.031	0.095	0.012	0.081	0.009	0.009	0.055	0.110	0.012	0.038	0.029	0.021	0.010	0.041	581	0.8988		
Polish	0.001	0.010	0.439	0.049	0.079	0.022	0.069	0.013	0.007	0.052	0.101	0.014	0.040	0.019	0.036	0.019	0.031	883	0.8991	VO	AN
Macedonian	0.005	0.000	0.450	0.045	0.075	0.045	0.040	0.030	0.010	0.020	0.080	0.020	0.050	0.015	0.040	0.030	0.045	200	0.9051	VO	AN
Spanish	0.027	0.009	0.440	0.081	0.070	0.019	0.058	0.010	0.013	0.019	0.083	0.018	0.062	0.013	0.013	0.017	0.049	3266	0.9083	VO	NA
Latvian	0.000	0.022	0.420	0.027	0.061	0.017	0.063	0.036	0.022	0.087	0.100	0.012	0.024	0.046	0.041	0.002	0.019	412	0.9178	VO	AN
Portuguese	0.064	0.001	0.439	0.048	0.068	0.033	0.063	0.012	0.009	0.017	0.065	0.030	0.061	0.022	0.018	0.020	0.029	1448	0.9179	VO	NA
Serbian-Croatian	0.001	0.033	0.411	0.051	0.093	0.018	0.053	0.017	0.002	0.025	0.101	0.004	0.050	0.032	0.033	0.010	0.068	838	0.9190	VO	AN
Czech	0.003	0.017	0.402	0.029	0.100	0.040	0.077	0.011	0.017	0.023	0.097	0.006	0.034	0.029	0.011	0.020	0.083	351	0.9224	VO	AN
Russian	0.002	0.018	0.411	0.042	0.078	0.027	0.065	0.014	0.011	0.039	0.104	0.022	0.037	0.020	0.018	0.013	0.080	1768	0.9313	VO	AN
Hungarian	0.003	0.008	0.392	0.049	0.101	0.025	0.087	0.008	0.005	0.033	0.074	0.014	0.068	0.019	0.052	0.011	0.052	367	0.9336	VO	AN
Chuvash	0.000	0.000	0.273	0.073	0.055	0.036	0.036	0.000	0.018	0.164	0.146	0.036	0.073	0.018	0.000	0.000	0.073	55	0.9419	OV	AN
Noghay	0.004	0.038	0.207	0.008	0.023	0.012	0.050	0.038	0.046	0.015	0.031	0.046	0.019	0.027	0.038	0.042	0.353	261	0.9445	OV	AN
Italian	0.008	0.030	0.407	0.033	0.086	0.033	0.084	0.014	0.019	0.017	0.052	0.026	0.076	0.014	0.017	0.018	0.066	3127	0.9491	VO	NA
Belorussian	0.000	0.020	0.391	0.060	0.089	0.023	0.051	0.023	0.020	0.031	0.091	0.066	0.023	0.017	0.037	0.003	0.054	350	0.9520	NDO	AN
Bulgarian	0.003	0.038	0.419	0.036	0.077	0.047	0.047	0.013	0.020	0.039	0.069	0.025	0.058	0.014	0.025	0.018	0.052	989	0.9530	VO	AN
Arabic (Moroccan)	0.206	0.020	0.281	0.074	0.047	0.007	0.042	0.005	0.010	0.007	0.040	0.099	0.048	0.000	0.006	0.025	0.083	836	0.9565	VO	NA
Ukrainian	0.003	0.035	0.384	0.043	0.081	0.029	0.084	0.017	0.009	0.058	0.098	0.006	0.049	0.017	0.026	0.009	0.058	346	0.9572	VO	AN
Romanian	0.002	0.020	0.369	0.040	0.104	0.035	0.051	0.012	0.010	0.029	0.071	0.024	0.079	0.009	0.039	0.019	0.090	1375	0.9677	VO	NA
Greek (Modern)	0.004	0.027	0.388	0.016	0.089	0.032	0.065	0.006	0.040	0.028	0.047	0.031	0.071	0.022	0.016	0.040	0.083	1162	0.9708	VO	AN
Tatar	0.000	0.009	0.307	0.040	0.075	0.026	0.092	0.009	0.022	0.070	0.105	0.031	0.057	0.009	0.018	0.000	0.132	228	0.9724	OV	AN
Turkish	0.013	0.048	0.309	0.007	0.089	0.028	0.043	0.013	0.037	0.019	0.030	0.063	0.056	0.015	0.019	0.019	0.192	463	0.9885	OV	AN
Ossetic	0.007	0.030	0.262	0.013	0.161	0.043	0.023	0.036	0.033	0.000	0.069	0.056	0.023	0.013	0.023	0.030	0.180	305	0.9900	OV	AN
Arabic (Egyptian)	0.218	0.041	0.157	0.004	0.088	0.053	0.061	0.006	0.033	0.010	0.016	0.029	0.045	0.020	0.010	0.006	0.206	491	0.9965	VO	NA
Avar	0.000	0.082	0.230	0.049	0.016	0.000	0.016	0.000	0.000	0.066	0.131	0.082	0.131	0.016	0.082	0.049	0.049	61	1.0037	OV	AN
Catalan	0.025	0.000	0.300	0.075	0.075	0.013	0.075	0.013	0.025	0.038	0.100	0.038	0.100	0.013	0.050	0.025	0.025	80	1.0135	VO	NA
Armenian (Eastern)	0.000	0.059	0.291	0.000	0.096	0.055	0.050	0.009	0.050	0.041	0.036	0.082	0.073	0.014	0.009	0.036	0.100	220	1.0207	NDO	AN
Dargwa	0.000	0.091	0.236	0.009	0.109	0.018	0.018	0.036	0.018	0.064	0.136	0.027	0.118	0.064	0.009	0.036	0.009	110	1.0255	OV	AN
Arabic (Modern Standard)	0.114	0.030	0.251	0.019	0.094	0.035	0.041	0.007	0.034	0.007	0.018	0.059	0.067	0.012	0.011	0.035	0.165	2914	1.0288	VO	NA
Chechen	0.000	0.011	0.258	0.007	0.032	0.018	0.125	0.029	0.061	0.039	0.104	0.029	0.050	0.057	0.032	0.014	0.136	279	1.0298	OV	AN
Karachay-Balkar	0.000	0.020	0.224	0.004	0.024	0.041	0.049	0.041	0.069	0.012	0.029	0.110	0.045	0.016	0.073	0.041	0.211	246	1.0306	OV	AN
Kumyk	0.000	0.036	0.241	0.009	0.080	0.018	0.098	0.027	0.063	0.063	0.054	0.027	0.027	0.054	0.018	0.009	0.179	112	1.0333		
Arabic (Syrian)	0.059	0.025	0.263	0.025	0.085	0.034	0.085	0.009	0.051	0.025	0.017	0.102	0.059	0.000	0.025	0.009	0.127	118	1.0375	VO	NA
Georgian	0.000	0.038	0.199	0.004	0.031	0.046	0.088	0.012	0.046	0.084	0.042	0.046	0.119	0.004	0.031	0.027	0.184	261	1.0407	OV	AN
Kurdish (Central)	0.023	0.084	0.169	0.006	0.140	0.051	0.034	0.011	0.028	0.006	0.017	0.152	0.051	0.034	0.028	0.011	0.163	178	1.0467	OV	NA
Persian	0.002	0.075	0.169	0.006	0.138	0.031	0.049	0.016	0.028	0.010	0.033	0.118	0.073	0.024	0.024	0.028	0.177	509	1.0483	OV	NA
Arabic (Iraqi)	0.076	0.092	0.172	0.002	0.132	0.054	0.035	0.009	0.050	0.028	0.014	0.073	0.052	0.009	0.021	0.012	0.168	424	1.0647	VO	NA
Adyghe (Abzakh)	0.000	0.043	0.256	0.000	0.046	0.036	0.062	0.036	0.072	0.020	0.085	0.030	0.072	0.023	0.053	0.049	0.121	305	1.0653	OV	NA
Azerbaijani	0.000	0.060	0.233	0.026	0.060	0.043	0.095	0.009	0.035	0.035	0.043	0.086	0.043	0.026	0.026	0.035	0.147	116	1.0743	OV	AN

Appendix 2 MtHg entropy and YHg entropy for 42 languages

Language	MtHG_ent	YHG_ent	OV-VO	AN-NA
Albanian	0.8841	0.6170	VO	NA
Arabic (Egyptian)	0.9965	0.5830	VO	NA
Arabic (Modern Standard)	1.0288	0.5710	VO	NA
Armenian (Eastern)	1.0207	0.7250	NOD	AN
Avar	1.0037	0.4730	OV	AN
Azerbaijani	1.0743	0.7910	OV	AN
Bashkir	0.8044	0.0950	OV	AN
Basque	0.8082	0.1680	OV	NA
Belorussian	0.9520	0.4890	NOD	AN
Bulgarian	0.9530	0.6990	VO	AN
Catalan	1.0135	0.4250	VO	NA
Chechen	1.0298	0.3730	OV	AN
Chuvash	0.9419	0.5720	OV	AN
Dutch	0.8620	0.4780	NOD	AN

Language	MtHG_ent	YHG_ent	OV-VO	AN-NA
English	0.8690	0.2540	VO	AN
Finnish	0.8719	0.3690	VO	AN
French	0.8955	0.4380	VO	NA
Georgian	1.0407	0.7030	OV	AN
German	0.8753	0.4480	NOD	AN
Greek (Modern)	0.9708	0.6990	VO	AN
Hungarian	0.9336	0.6030	VO	AN
Icelandic	0.8686	0.2800	VO	AN
Irish	0.8525	0.0000	VO	NA
Italian	0.9491	0.5670	VO	NA
Karachay-Balkar	1.0306	0.6300	OV	AN
Kurdish (Central)	1.0467	0.6590	OV	NA
Latvian	0.9178	0.3240	VO	AN
Ossetic	0.9900	0.5350	OV	AN

Language	MtHG_ent	YHG_ent	OV-VO	AN-NA
Persian	1.0483	0.6800	OV	NA
Polish	0.8991	0.3890	VO	AN
Romanian	0.9677	0.6510	VO	NA
Russian	0.9313	0.5390	VO	AN
Saami (Northern)	0.4625	0.4630	VO	AN
Serbian-Croatian	0.9190	0.5710	VO	AN
Slovene	0.8934	0.4310	VO	AN
Spanish	0.9083	0.4960	VO	NA
Swedish	0.8686	0.4700	VO	AN
Tatar	0.9724	0.5120	OV	AN
Turkish	0.9885	0.7290	OV	AN
Udmurt	0.8461	0.1920	OV	AN
Ukrainian	0.9572	0.5260	VO	AN
Welsh	0.6102	0.0620	VO	NA