

Relation between Word Order Characteristics and Suicide/Homicide Rates (3)

語順特徴と自殺率／他殺率との関係(その 3)

Terumasa EHARA

江原暉将

Yamanashi Eiwa College

山梨英和大学

<http://www.yamanashi-eiwa.ac.jp/~eharate/>

1 Introduction

The previous papers [Ehara, 2010, 2011] showed quantitative relations between the word order characteristics and suicide/homicide rates. Our study purpose is to clarify relations between syntactic structures of languages, especially word order structures, and people's thinking pattern who use that language as a native language [Ehara, 1995, 2007]. In this paper, we add non-linguistic features: economic feature and climate features in addition to linguistic features to analyze the relation.

Death is the most important event for all human beings. Suicide and homicide are abnormal death. Then, we suppose that people's thinking pattern affects the suicide and homicide. We can measure them quantitatively with suicide rate and homicide rate. So, we use them as the measure of thinking pattern.

2 Data

Data for the word order characteristics (features) are obtained from the WALS database [Dryer, 2005]. The number of languages analyzed in this article is 1473. The following thirteen word order features are considered in our analysis.

- (1) Order of Subject(S) and Verb(V)
- (2) Order of Object(O) and Verb
- (3) Order of Oblique(X) and Verb
- (4) Order of Adposition(Ad) and Noun phrase(N)
- (5) Order of Genitive(G) and Noun
- (6) Order of Adjective(A) and Noun
- (7) Order of Demonstrative(Dm) and Noun
- (8) Order of Numeral(Nm) and Noun
- (9) Order of Relative clause(R) and Noun
- (10) Order of Degree word(Dg) and Adjective
- (11) Position of Polar question particles
- (12) Position of Interrogative phrases in content questions

- (13) Order of Adverbial subordinator(As) and Clause(C)

We define feature value "+" if the order is same as Japanese and "-" if the order is opposite of the "+". The feature value is "0" if the order is other than "+" or "-" and "." if the feature value is not described in WALS database. Table 1 shows all feature values for the thirteen features.

Table 1: Word order feature values

| No. | + | - |
|-----|-------------|---------|
| 1 | SV | VS |
| 2 | OV | VO |
| 3 | XV | VX |
| 4 | NAp | ApN |
| 5 | GN | NG |
| 6 | AN | NA |
| 7 | DmN | NDm |
| 8 | NmN | NNm |
| 9 | RN | NR |
| 10 | DgA | ADg |
| 11 | Final | Initial |
| 12 | Not initial | Initial |
| 13 | CAs | AsC |

Suicide rate and homicide rate are obtained from the WHO's "mortality and burden of disease estimates for WHO member states in 2004" [WHO, 2009]. From this database, we can obtain suicide and homicide rate for 192 countries or regions of the world. We use log10 values of these rates instead of rates themselves.

Language names spoken in countries and regions are obtained from Nations Online [Nationsonline, 2006]. This table includes 218 country names with official or national language names. We use the firstly listed language name in the table as the language name used in the country.

Combining the above three databases, we got the data for 178 countries which include

73 languages.

As economic feature, we use GDP per capita (simply, GDP). Feature values of GDP are obtained from a data access system to UN databases (UNdata)[UN, 2011]¹. The data which are measured by US dollars are converted by log10 to obtain the normality of the distribution.

For climate features, we use average annual temperature (degrees centigrade, TMP) and average annual precipitation (cm, PRC) at capital cities of countries. If these data are not available for some capital city, we use the data at an observation point nearest to that capital city. Climate feature values are obtained from [NAOJ, 2012], [World Climate, 2005] and other internet sites. Table 2 shows the number of climate data for the three data sauces.

Table 2: Number of climate data

| Data sauces | Temperature | Precipitation |
|---------------|-------------|---------------|
| NAOJ | 125 | 129 |
| World climate | 32 | 35 |
| Other sauces | 21 | 14 |

Log10 of suicide rate (S-rate) and log10 of homicide rate (H-rate) for 178 countries are distributed as Figure 1. Sample mean and sample standard deviation of data are listed in Table 3.

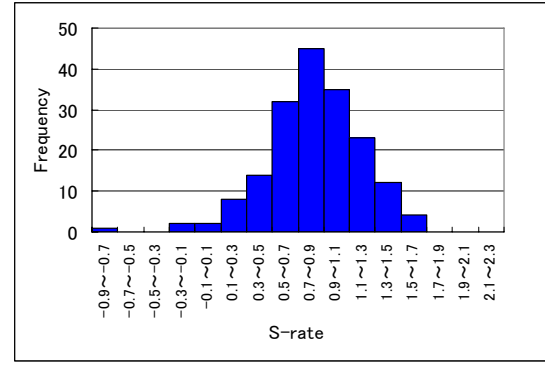
3 Analysis and results

Our analysis consists of three steps. Step 1 makes multiple regression analysis for S-rate and H-rate. Explanatory variables are GDP, TMP and PRC. We use S-rate and H-rate as the criterion variables instead of SH-ratio which was used in previous papers. Residuals of the step 1 are considered as unexplainable data by the explanatory variables. So, we make a t-test to residuals by word order characteristics in step 2 and step 3. In step 2, we merge residual data by language names. In this merging process, we make weighting by population. It is the reason why to use S-rate and H-rate instead of SH-ratio. Step 3 makes t-test by word order features which was similar to the previous paper [Ehara, 2011].

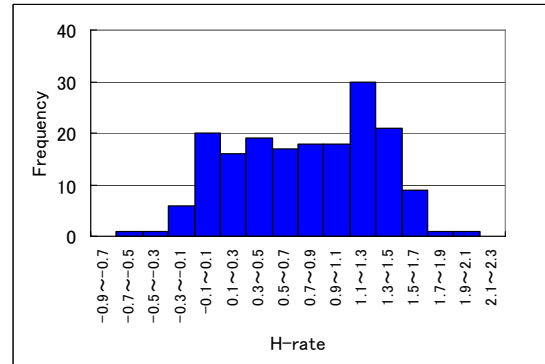
Step1: Multiple regression analysis

Criterion variables are S-rate and H-rate. Explanatory variables are GDP, TMP and PRC.

¹ GDP per capita for Niue is not included in UNdata. We got the data from <http://www.indexmundi.com/g/g.aspx?c=ne&v=67>.



(a) S-rate



(b) H-rate

Figure 1: Distribution of Data

Table 3: Statistics of S-rate and H-rate

| | Sample mean | Sample S.D. | S.D. of residuals |
|--------|-------------|-------------|-------------------|
| S-rate | 0.8276 | 0.3632 | 0.3330 |
| H-rate | 0.7480 | 0.5426 | 0.4316 |

Summary of the results is shown in Table 4. Contribution ratio for S-rate and H-rate are 0.1592 and 0.3673, respectively. Looking at Table 4, S-rate has higher correlations with TMP (negative) and PRC (positive) than with GDP. In contrast, H-rate has higher correlation with GDP (negative) than TMP and PRC. Standard deviation of residuals for S-rate and H-rate are listed in Table 3. Figure 2 shows scatter diagram of predicted values and observed values of this regression.

Step2: Merging

We merge S-rate and H-rate residuals by language names. Merged residuals for S-rate for language l ($RS - rate(l)$) is defined by

$$RS - rate(l) = \log_{10} \left(\frac{\sum_{c \in C(l)} 10^{RS - rate(c)} \times pop(c)}{\sum_{c \in C(l)} pop(c)} \right)$$

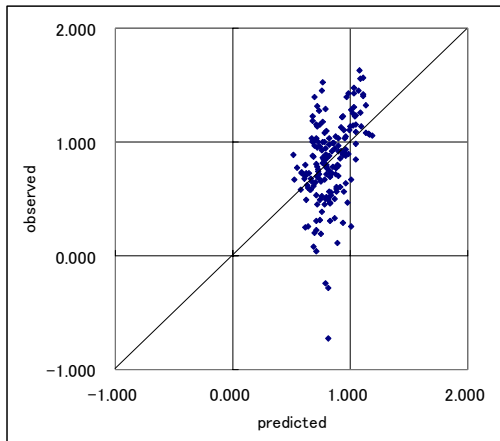
where $C(l)$ is the set of countries which use

Table 4: Results of multiple regression analysis
(a) S-rate

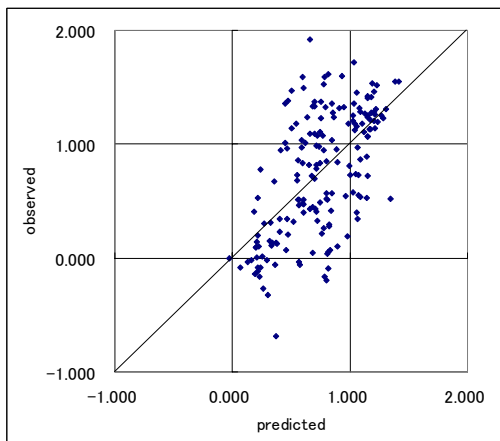
| For S-rate | Partial regression coefficient | standardized partial regression coefficient | Correlation coefficient | Partial correlation coefficient |
|------------|--------------------------------|---|-------------------------|---------------------------------|
| GDP | 0.0228 | 0.0430 | 0.1833 | 0.0422 |
| TMP | -0.0215 | -0.4213 | -0.3547 | -0.3574 |
| PRC | 0.0008 | 0.1989 | 0.0093 | 0.1923 |
| Intercept | 1.0856 | 0.0000 | | |

(b) H-rate

| For H-rate | Partial regression coefficient | Standardized partial regression coefficient | Correlation coefficient | Partial correlation coefficient |
|------------|--------------------------------|---|-------------------------|---------------------------------|
| GDP | -0.4517 | -0.5704 | -0.6016 | -0.5425 |
| TMP | 0.0068 | 0.0885 | 0.3214 | 0.0923 |
| PRC | -0.0002 | -0.0339 | 0.1257 | -0.0385 |
| Intercept | 2.2610 | 0.0000 | | |



(a) S-rate



(b) H-rate

Figure 2: Predicted value and observed value

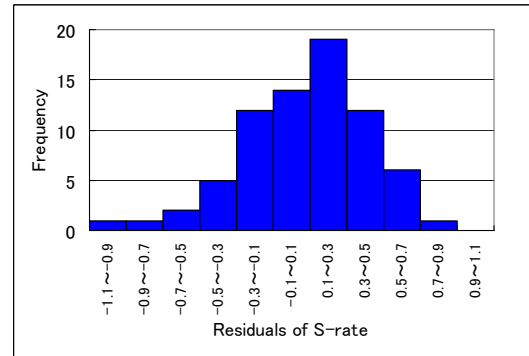
the language l , $pop(c)$ is population of the country c and $RS-rate(c)$ is the residuals of S-rate of the country c . Merged residuals for H-rate ($RH-rate(l)$) for language l is defined similarly. From this merging process, we get RS-rate and RH-rate data for 73 languages. Frequency distribution of these data is shown in Figure 3.

Step3: T-test

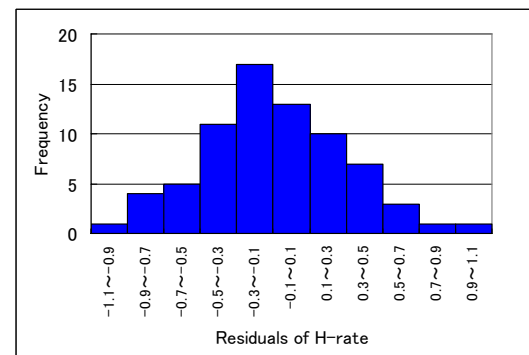
For each word order features, we divide RS-rate and RH-rate data to the plus group and the minus group and apply the t-test between two groups. The result is shown in Table 5. We do not make discussion about features which have data counts of + or - are less than 10 because the number of languages are too small to make discussion. These feature numbers are 1, 3, 8, 10 and 13.

For RS-rate, there is no significant (with significance level 5%) feature. Feature 6 has largest t-value but it is not significant.

For RH-rate, feature 2 and 12 are significant.



(a) RS-rate



(b) RH-rate

Figure 3: Distribution of residuals

4 Conclusion

We examine the relation between economic, climate and linguistic features and suicide / homicide rates². Firstly, multiple regression analysis is applied in which GDP, tempera-

² Used data will be opened at my web site.

ture and precipitation are used as explanatory variables. Next, t-test is applied to residuals of the above multiple regression using linguistic features.

From the results, we can conclude that:

- (a) Suicide rate has higher correlations with temperature (negative) and precipitation (positive) than with GDP.
- (b) Homicide rate has higher correlation with GDP (negative) than with temperature and precipitation.
- (c) OV word order group has lower homicide rate and VO word order group has higher homicide rate.
- (d) The group, that interrogative phrase of content question is located at non-initial position, has lower homicide rate and the group, that interrogative phrase of content question is located at initial position, has higher homicide rate.

Other linguistic features may affect suicide and homicide rates. Study using these features is remained as a future work.

References

- [Dryer, 2005] Dryer, Matthew S.: Word Order, The World Atlas of Language Structures, Chapter F, pp.330-397, Oxford University Press, 2005.
<http://wals.info/>
- [Ehara, 1995] EHARA, Terumasa : Relation among Word Order Parameters Analyzed by Multi-Dimensional Scaling, Proceedings of The first Annual Meeting of The Association for Natural Language Processing, pp.173-176, Mar., 1995 (in Japanese).
- [Ehara, 2007] EHARA, Terumasa : Word Order Characteristics Analyzed by Multi Dimensional Scaling, Proceedings of The 13th Annual Meeting of The Association for Natural Language Processing, A1-3, Mar., 2007.
- [Ehara, 2010] EHARA, Terumasa : Relation between the Word Order Characteristics and Suicide/Homicide Rates, Proceedings of The 16th Annual Meeting of The Association for Natural Language Processing, E4-2, pp.956-959, Mar., 2010.
- [Ehara, 2011] EHARA, Terumasa : Relation between the Word Order Characteristics and Suicide/Homicide Rates (2), Proceedings of The 17th Annual Meeting of The Association for Natural Language Processing, F4-6, pp.1037-1040, Mar., 2011.
- [NAOJ, 2012] National Astronomical Observatory of Japan(Ed): Chronological Scientific Tables, Maruzen Co., Ltd., pp.269-317, 2012.
- [Nationsonline, 2006] One World - Nations Online : Official and National Languages of the World by Continent, 2006 version.
<http://www.nationsonline.org/oneworld/languages.htm>
- [UN, 2011] United Nations Statistics Division: Per

capita GDP at current prices - US dollars, A data access system to UN databases (UNdata), 2011.

<http://data.un.org/Default.aspx>

[WHO, 2009] World Health Organization : Mortality and burden of disease estimates for WHO member states in 2004.

http://www.who.int/entity/healthinfo/global_burden_disease/gbddeathdalycountryestimates2004.xls

[World climate, 2005] World climate: WorldClimate, v.271, 2005.

<http://www.worldclimate.com/>

Table 5: T-test results for the thirteen features

(a) RS-rate

| feature number | data counts | | sample mean | | sample standard deviation | | t-value |
|----------------|-------------|----|-------------|--------|---------------------------|-------|---------|
| | + | - | + | - | + | - | |
| 1 | 56 | 6 | 0.121 | -0.220 | 0.317 | 0.216 | 3.484 |
| 2 | 20 | 46 | 0.042 | 0.056 | 0.406 | 0.320 | -0.139 |
| 3 | 3 | 17 | 0.113 | 0.098 | 0.357 | 0.197 | 0.071 |
| 4 | 17 | 46 | 0.143 | 0.055 | 0.339 | 0.317 | 0.925 |
| 5 | 26 | 34 | 0.091 | -0.013 | 0.337 | 0.344 | 1.174 |
| 6 | 40 | 31 | 0.129 | 0.001 | 0.323 | 0.352 | 1.567 |
| 7 | 50 | 11 | 0.089 | 0.061 | 0.282 | 0.299 | 0.293 |
| 8 | 54 | 8 | 0.086 | 0.116 | 0.319 | 0.278 | -0.282 |
| 9 | 10 | 47 | 0.121 | 0.062 | 0.264 | 0.345 | 0.601 |
| 10 | 37 | 6 | 0.174 | 0.010 | 0.263 | 0.386 | 1.001 |
| 11 | 13 | 19 | -0.079 | 0.082 | 0.480 | 0.308 | -1.065 |
| 12 | 29 | 20 | 0.097 | 0.031 | 0.364 | 0.348 | 0.648 |
| 13 | 5 | 49 | 0.127 | 0.064 | 0.220 | 0.356 | 0.570 |

(b) RH-rate

| feature number | data counts | | sample mean | | sample standard deviation | | t-value |
|----------------|-------------|----|-------------|--------|---------------------------|-------|---------|
| | + | - | + | - | + | - | |
| 1 | 56 | 6 | -0.071 | -0.358 | 0.351 | 0.487 | 1.407 |
| 2 | 20 | 46 | -0.238 | -0.021 | 0.287 | 0.413 | -2.448 |
| 3 | 3 | 17 | -0.343 | 0.019 | 0.179 | 0.347 | -2.712 |
| 4 | 17 | 46 | -0.075 | -0.085 | 0.269 | 0.408 | 0.112 |
| 5 | 26 | 34 | -0.101 | -0.109 | 0.321 | 0.443 | 0.075 |
| 6 | 40 | 31 | -0.073 | -0.133 | 0.341 | 0.431 | 0.642 |
| 7 | 50 | 11 | -0.056 | -0.183 | 0.374 | 0.411 | 0.947 |
| 8 | 54 | 8 | -0.108 | -0.181 | 0.361 | 0.445 | 0.440 |
| 9 | 10 | 47 | -0.201 | -0.078 | 0.248 | 0.440 | -1.210 |
| 10 | 37 | 6 | -0.062 | -0.230 | 0.401 | 0.367 | 1.025 |
| 11 | 13 | 19 | -0.233 | -0.043 | 0.274 | 0.461 | -1.461 |
| 12 | 29 | 20 | -0.279 | 0.108 | 0.320 | 0.399 | -3.611 |
| 13 | 5 | 49 | -0.147 | -0.131 | 0.233 | 0.413 | -0.139 |