Relation between Word Order Characteristics and Suicide / Homicide Rates (4)
語順特徴と自殺率／他殺率との関係(その 4)

Terumasa EHARA
江原暉将

Yamanashi Eiwa College
山梨英和大学
http://www.yamanashi-eiwa.ac.jp/~eharate/

1 Introduction

In previous papers [Ehara, 2010, 2011, 2012], we have investigated quantitative relations between the word order characteristics and suicide rate / homicide rate. The purpose of our study is to clarify relations between syntactic structures of a language, especially word order structures, and people's thinking pattern who use it as a native language.

In the last paper [Ehara, 2012], we added non-linguistic features: economic feature (GDP per capita) and climate features (average annual temperature and average annual precipitation) in addition to linguistic features to analyze the relation. The analysis in that paper consists of three steps. First step is multiple regressions in which economic and climate features are explanatory variables and suicide rate and homicide rate are criterion variables. The first step was done by country based data. In the second step, residuals of the first step were merged by language names so that we can get language based data. In the third step, we made t-tests for distributions of merged residuals by word order features. The reason why these three steps were made was that we could not merge country based data to language based data at that time.

In this paper, we merge country based data to language based data using weighted average by population of countries. So, we can make single step analysis in this paper.

2 Data

Data for the word order characteristics (features) are obtained from the WALS database [Dryer, 2005]. In this paper, we use only two features:
(1) Order of Object (O) and Verb (V)
(2) Order of Adjective (A) and Noun (N).
The reason only to use these two features is that we want to use as many languages as possible of which word order features are all defined. As described later, we can define both two features to 64 languages within 73 languages. Other reason is that these two features are major feature expressing the word order structure of languages [Ehara, 1995, 2007].

We define feature value "1" if the order is same as Japanese and "－1" if the order is opposite of "1". For feature-1 (Fea-1), OV corresponds to 1 and VO corresponds to －1. For feature-2 (Fea-2), AN corresponds to 1 and NA corresponds to －1.

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].

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

Combining the above three databases, we get the data for 178 countries which include 73 languages. Language names used in more than two countries are listed in Table 1.

Language names are listed in Figure 1. which are sorted by the word order types. In our analysis, we only use 64 languages excluding 9 languages in Figure 1 (e).

How to get GDP per capita data, average annual temperature data and average annual precipitation data was explained in [Ehara, 2012][1].

[1] We use three variables: GDP which is log10 of GDP per capita measured by U.S. dollar, TMP which is average annual temperature measured by ℃ and PRC which is average annual precipitation measured by cm.

Table 1: Languages used in more than one country

Language	No. of countries
English	42
Spanish	20
French	18
Arabic (Modern Standard)	11
Portuguese	7
Arabic (Gulf)	4
Dutch	3
German	3
Arabic (Moroccan)	2
Greek (Modern)	2
Italian	2
Korean	2
Mandarin	2

Albanian, Arabic(Egyptian), Arabic(Gulf), Arabic(Iraqi), Arabic(Modern Standard), Arabic(Moroccan), Arabic(Syrian), Catalan, French, Hebrew(Modern), Indonesian, Irish, Italian, Khmer, Kinyarwanda, Lao, Nauruan, Portuguese, Romanian, Samoan, Sesotho, Spanish, Swahili, Thai, Tongan, Vietnamese

(a) VO and NA type (26 languages)

Bulgarian, Czech, Danish, English, Estonian, Finnish, Greek(Modern), Icelandic, Latvian, Lithuanian, Macedonian, Mandarin, Norwegian, Polish, Russian, Serbian, Slovene, Swedish, Ukrainian

(b) VO and AN type (19 languages)

Burmese, Motu, Persian, Somali, Tajik

(c) OV and NA type (5 languages)

Amharic, Azerbaijani, Georgian, Hindi, Japanese, Khalkha, Korean, Nepali, Pashto, Sinhala, Tigrinya, Turkish, Turkmen, Urdu, Uzbek

(d) OV and AN type (15 languages)

Belorussian, Dutch, German, Hungarian, Tagalog, Rundi, Armenian(Eastern), Dhivehi

(e) Other word order types (9 languages)

Figure 1: Language names sorted by word order types

We merge suicide rate, homicide rate, GDP per capita data, average annual temperature data and average annual precipitation data from country based to language based. We use weighted average by population of countries as merged data. For example, if language L is used in countries $C_1, C_2, \cdots, C_n$, average annual temperature T_L for language L is

$$T_L = (\sum_{i=1}^{n} P_{C_i} \times T_{C_i}) / \sum_{i=1}^{n} P_{C_i} \quad (1)$$

where T_{C_i} is average annual temperature for country C_i and P_{C_i} is population of the country C_i. It is natural to use (1) because climate condition T_{C_i} affects all peoples living in the country C_i.

We define two variables: S-rate which is log10 of suicide rate and H-rate which is log10 of homicide rate. Appendix shows the distribution of S-rates and H-rates for 64 languages.

3 Analysis and results

We make multiple regression analysis. Criterion variables are S-rate and H-rate. Explanatory variables are GDP, TMP, PRC, Fea-1 and Fea-2. Contribution ratio for S-rate and H-rate are 0.2779 and 0.3154, respectively.

Summary of the results is shown in Table 2.

Table 2: Results of multiple regression analysis

(a) S-rate

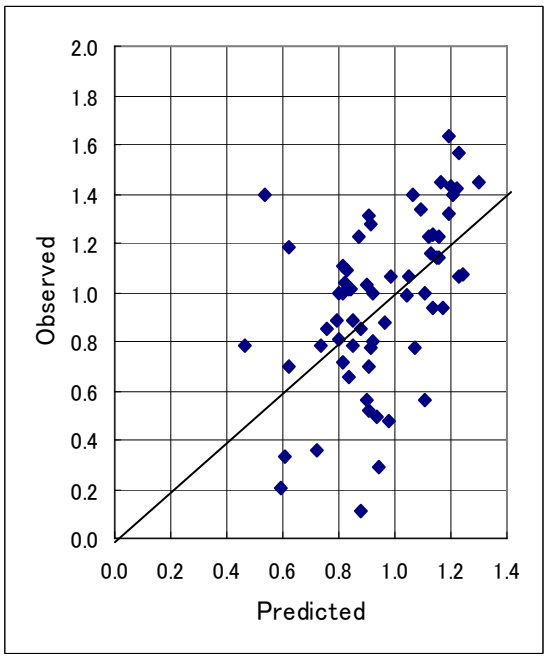
S-rate	Partial regression coefficient	Standardized partial regression coefficient	T-test value	Correlation coefficient	Partial correlation coefficient
GDP	−0.045	−0.079	−0.512	0.213	−0.067
TMP	−0.012	−0.228	−1.270	−0.267	−0.163
PRC	0.002	0.385	2.865	0.122	0.349
Fea-1	−0.095	−0.237	−1.671	−0.084	−0.213
Fea-2	0.168	0.454	2.951	0.368	0.359
Intercept	1.054	0.000	2.769		

(b) H-rate

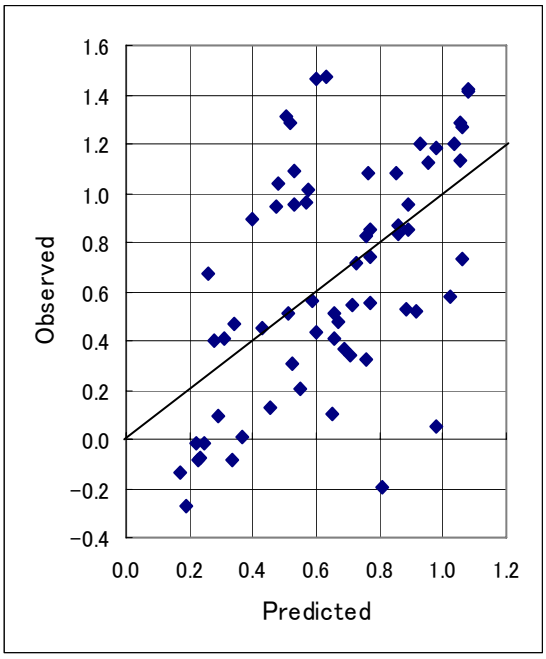
H-rate	Partial regression coefficient	Standardized partial regression coefficient	T-test value	Correlation coefficient	Partial correlation coefficient
GDP	−0.449	−0.631	−4.187	−0.499	−0.479
TMP	−0.001	−0.020	−0.115	0.323	−0.015
PRC	0.001	0.203	1.552	0.226	0.198
Fea-1	−0.127	−0.250	−1.808	0.073	−0.229
Fea-2	0.064	0.137	0.914	−0.150	0.118
Intercept	2.117	0.000	4.502		

Looking at Table 2, S-rate has higher correlations with Fea-2 (positive) and PRC (positive) than with GDP, TMP and Fea-1. Fea-2 and PRC has significant non-zero regression coefficient with 5% significance level. In contrast, H-rate has higher correlation with GDP (negative) and Fea-1 (negative) than TMP, PRC and Fea-2. However, only GDP has significant non-zero regression coefficient with 5% significance level. Table 3 shows correlation matrices. Figure 2 shows scatter diagram of predicted values and ob-

served values of the regressions.



(a) S-rate



(b) H-rate

Figure 2: Predicted value and observed value

Table 3: Correlation matrices of multiple regression analysis

(a) S-rate

	S-rate	GDP	TMP	PRC	Fea-1	Fea-2
S-rate	1.000	0.213	−0.267	0.122	−0.084	0.368
GDP	0.213	1.000	−0.532	−0.125	−0.450	0.246
TMP	−0.267	−0.532	1.000	0.533	0.082	−0.589
PRC	0.122	−0.125	0.533	1.000	−0.003	−0.336
Fea-1	−0.084	−0.450	0.082	−0.003	1.000	0.303
Fea-2	0.368	0.246	−0.589	−0.336	0.303	1.000

(b) H-rate

	H-rate	GDP	TMP	PRC	Fea-1	Fea-2
H-rate	1.000	−0.499	0.323	0.226	0.073	−0.150
GDP	−0.499	1.000	−0.532	−0.125	−0.450	0.246
TMP	0.323	−0.532	1.000	0.533	0.082	−0.589
PRC	0.226	−0.125	0.533	1.000	−0.003	−0.336
Fea-1	0.073	−0.450	0.082	−0.003	1.000	0.303
Fea-2	−0.150	0.246	−0.589	−0.336	0.303	1.000

4 Conclusion

We examine the relation between economic, climate and linguistic features and suicide rate (S-rate) / homicide rate (H-rate)[2]. We make multiple regression analysis. Explanatory variables are GDP per capita (GDP), average annual temperature (TMP), average annual precipitation (PRC) and two word order features. Two word order features are order of Verb and Object noun (Fea-1: OV is +1 and VO is −1) and order of Adjective and Noun (Fea-2: AN is +1 and NA is −1). S-rat, H-rate, GDP, TMP and PRC data are given in country based. We merge them to language based data using weighted average by population.

From the results, we can conclude that:

(a) S-rate has higher correlations with Fea-2 (positive) and PRC (positive) than with GDP, TMP and Fea-1. Fea-2 and PRC have significantly non zero regression coefficient with 5% significant level. AN word order has higher suicide rate than NA word order.

(b) H-rate has higher correlation with GDP (negative) and Fea-1 (negative) than with TMP, PRC and Fea-2. However significant variable is only GDP with 5% significant level. VO word order has higher homicide rate than OV word order, but it is not significant.

Other linguistic features and non-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

[2] Used data are opened at my web site.

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.

[Ehara, 2012] EHARA, Terumasa : Relation between the Word Order Characteristics and Suicide/Homicide Rates (3), Proceedings of The 18th Annual Meeting of The Association for Natural Language Processing, B1-5, pp.54-57, Mar., 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

[WHO, 2009] World Health Organization : Mortality and burden of disease estimates for WHO member states in 2004. http://www.who.int/entity/healthinfo/global_buden_disease/gbddeathdalycountryestimates2004.xls

Appendix Distribution of S-rate and H-rate for 63 languages

Language names are expressed in WALS code; Red diamond: NA language; Blue diamond: AN language.