Structural Distance, Lexical Distance and Speaker’s Genetic Distance of Languages
各種言語の構造的距離、語彙的距離および話者の遺伝的距離

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

It is true that any human-being can use any language when he/she is exposed in the language environment at his/her language acquisition age. This fact, however, does not mean that there are no relation between a language and a population of the speakers of the language. For example, Cavalli-Sforza et al. (1992) shows that there are relations between language groups and populations.

The purpose of our research is to clarify the relation between linguistic (structural and lexical) distances and genetic distance of speakers of languages. By the author’s knowledge, previous studies for the relation between languages and populations do not use language names but family names or group names. We calculate linguistic distances using language names and compare genetic distance of the speakers of these languages.

2 Data

Structural data of languages are obtained from WALS database (Dryer and Haspelmath, 2013). This database contains feature values for 192 structural features of 2,679 languages. We use only 142 main features for our analysis that are WALS feature number 1 to 142 and suffixed “A”.

We also select 1049 languages from the original database. We use languages which have more than 29 feature values or more than 8 word order feature values[1] to ensure a reliable analysis. We do not use sign languages, pidgins and creoles.

Lexical data of languages are obtained from ASJP database (Wichmann et al., 2013). This database contains 100 words of Swadesh list for 6,895 dialects[2] with ISO language code if any. We select 3,345 dialects which have the ISO code to which the WALS code corresponds. The number of language names in this selected data are 2,047.

Genetic data of populations are obtained from Wikipedia’s page of “Y-chromosome haplogroups by populations” (Wikimedia Foundation, 2015). Merging the genetic data from European, Near East’s, North African, Sub-Saharan African, Caucasus’, South Asian, East and Southeast Asian, North Asian and Oceanian populations and also the data from indigenous peoples of the Americas, we obtain relative frequency data of Y-chromosome haplogroups (YHg) for 505 populations. Using a population name, we estimate a language name which is used by the population. For example, we estimates that the population Evenks use the language Evenki. The data that we cannot estimate the language name are discarded. For example, the datum from India including Indo-European, Dravidian, Austro-Asiatic and Sino-Tibetan speakers are discarded. As the result, we can get the data from 452 populations. The number of language names are 196 in this selected data. Some languages are used by several populations. For example, Japanese is used by the populations: Japan, Japan (Kantō) and Western Japan. In the Wikipedia pages, the granularity of used YHg is different page by page. We adjust the granularity to the coarsest 20 groups from A to T (Karafet et al., 2008).

3 Dissimilarity Metrics

Structural dissimilarity of languages is measured by the relative Hamming distance (Ehara, 2009). For languages l_i and l_k, structural distance $d_s(l_i, l_k)$ is determined as the ratio of different feature values of l_i and l_k for all features of which feature values are

[1] There are 17 word order features that are 81A to 97A in our analysis.

[2] ASJP’s term “name” should be considered as “dialect name”. For example, there are four data named JAPANESE, JAPANESE_2, TOKYO_JAPANESE and JAPANESE_KYOTO in ASJP. They are all Japanese. So we use the term “dialect name” instead of the term “name” in distinction from the term “language name”.

defined both l_i and l_k. Explicitly

$$d_s(l_i, l_k) = \frac{\sum_{f \in F_{i,k}} d_f(l_i, l_k)}{N}$$

where $F_{i,k}$ is the set of features of which feature values are defined both l_i and l_k and N is the number of elements of $F_{i,k}$ and

$$d_f(l_i, l_k) = \begin{cases} 0 & if\ f(l_i) = f(l_k) \\ 1 & else \end{cases}$$

where $f(l)$ is the feature value for l.
For all language pairs, $0 \le d_s(l_i, l_k) \le 1$. If all feature values of l_i and l_k are same, $d_s(l_i, l_k) = 0$ and if all feature values of l_i and l_k are different, $d_s(l_i, l_k) = 1$.

Lexical dissimilarity of dialects d_i and d_k is measured by the mean value of lexical dissimilarities of 100 word pairs of d_i and d_k. Lexical dissimilarity $d_l(w_r, w_r')$ of words w_r and w_r' $(1 \le \mathrm{r} \le 100)$ is relative Levenshtein distance (edit distance), which is calculated by character base[3]. For all dialects d_i and d_k, lexical dissimilarity $d_l(d_i, d_k)$ is calculated by

$$d_l(d_i, d_k) = \frac{1}{N}\sum_{r=1}^{N} d_l(w_r, w_r')$$

where N is the number of both w_r and w_r' are defined. For all word pairs, $0 \le d_l(w_r, w_r') \le 1$ and for all dialect pairs, $0 \le d_l(d_i, d_k) \le 1$. Lexical dissimilarity of languages l_i and l_k is the minimum value of dissimilarities of the members of dialect sets D_i for l_i and D_k for l_k :

$$d_l(l_i, l_k) = \min_{d_m \in D_i,\ d_n \in D_k} d_l(d_m, d_n)$$

Genetic dissimilarity of populations is measured by the distance of probability function. Defining p_r^i is relative frequency of rth YHg $(1 \le \mathrm{r} \le 20)$ for population e_i and p_r^k is relative frequency of rth YHg $(1 \le \mathrm{r} \le 20)$ for population e_k, genetic dissimilarity $d_e(e_i, e_k)$ of populations e_i and e_k is calculated by the Nei's minimum genetic distance (Nei and Roychoudhury, 1974):

$$d_e(e_i, e_k) = \frac{\sum_{r=1}^{N} {\mathrm{p}_r^i}^2 + \sum_{r=1}^{N} {\mathrm{p}_r^k}^2}{2} - \sum_{r=1}^{N} \mathrm{p}_r^i \mathrm{p}_r^k.$$

where N is the number of YHg's (N=20). Genetic dissimilarity of languages l_i and l_k is the minimum value of dissimilarities of the members of population sets E_i for l_i and E_k for l_k :

$$d_e(l_i, l_k) = \min_{e_m \in E_i,\ e_n \in E_k} d_e(e_m, e_n).$$

4 Statistical Analysis and Results

Using three dissimilarity metrics, we can get three dissimilarity matrices for languages: structural dissimilarity matrix D_s (1049 × 1049), lexical dissimilarity matrix D_l (2047 × 2047) and genetic dissimilarity matrix D_e (196 × 196). We conduct multi-dimensional scaling (MDS) for these matrices by the Torgerson's method.

As the results, Eigen values greater or equal to 0.5 are used to make the configuration spaces. Dimension of the configuration spaces and cumulative contribution ratios of the adopted Eigen values are listed in Table 1.

Table 1: Results of MDS

	Dimension	Cum. cont. ratio
D_s	254	0.66
D_l	754	0.79
D_e	14	0.83

From the Euclidian distances of these configuration spaces, we can get three kinds of distances for languages: structural distance $\bar{d}_s(l_i, l_k)$, lexical distance $\bar{d}_l(l_i, l_k)$ and genetic distance $\bar{d}_e(l_i, l_k)$ for languages l_i and l_k.

Frequency distribution of distances by $\bar{d}_s(l_i, l_k)$, $\bar{d}_l(l_i, l_k)$ and $\bar{d}_e(l_i, l_k)$ $(i < k)$ are shown in Figure 1.

Pearson's correlation coefficients are calculated using two of the three distance data. Number of languages in this calculation are 167. Number of data points are 13,861. The results are shown in Table 2. They have weak correlations.

5 Around Japanese

Next, we restrict data such that l_i is fixed to Japanese. Number of data points in this restricted data against Japanese is 166. Correlation coefficients using this restricted data are shown in Table 3. Correlation coefficients between structural – lexical data and lexical – genetic data is almost zero. On the other hand, structural – genetic data also

[3] Relative Levenshtein distance is defined by the character based Levenshtein distance divided by the number of characters of the longer word of the two (Serva, 2009). Vowel nasalization, two juxtaposed consonants, three juxtaposed consonants and glottalized consonant are considered as one character (Holman, 2014). If w_r or w_r' are not defined, we set $d_l(w_r, w_r') = 0$. If rth word of d_i and d_k have more than one word (ex. JAPANESE_2 has three words: anata, kimi, omae for “you”), $d_l(w_r, w_r')$ is mininum value of all combination of rth word pairs of d_i and d_k.

has weak correlation as in Table 2. Scattering graph of structural – genetic distance data is shown in Figure 2.

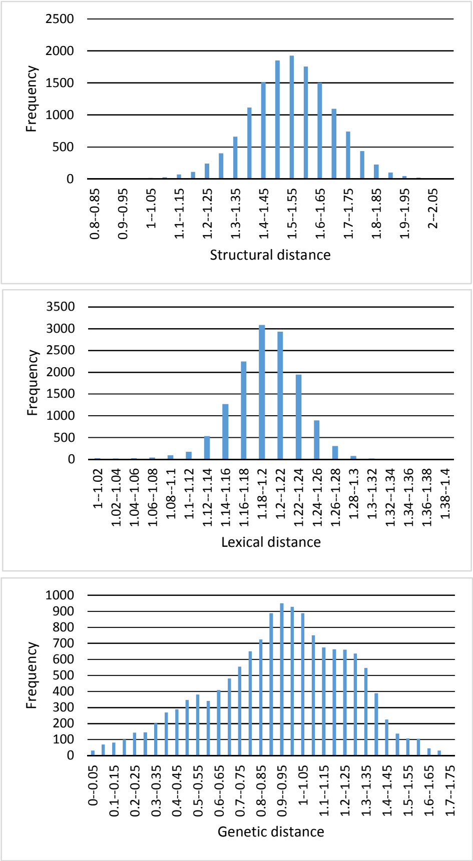


Figure 1: Frequency distribution of distances

Table 2: Correlation coefficients

	Corr. coef.
Structural – Lexical	0.217
Lexical – Genetic	0.201
Structural – Genetic	0.231

Table 3: Correlation coefficients using restricted data against Japanese

	Corr. coef.
Structural – Lexical	0.014
Lexical – Genetic	0.081
Structural – Genetic	0.229

Top 20 languages close to Japanese are listed in Table 4. The nearest languages by structural distance, lexical distance and genetic distance are Korean, Shuri (Ryukuu) and Tibetan, respectively.

Several lexically close languages are strange. Warekena is a language of Arawakan family spoken in Brazil, Colombia and Venezuela.

Some structurally close languages are overlapped to genetically close languages. They are Mandarin, Korean and Garo. Garo belongs to Sino-Tibetan family Tibeto-Burman subfamily Bodo-Garo Genus. It is spoken in the north east area of India.

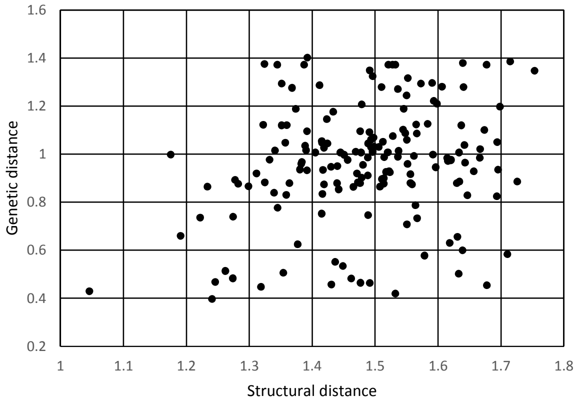


Figure 2: Scattering graph of structural distance and genetic distance data against Japanese

6 Conclusion and Future Works

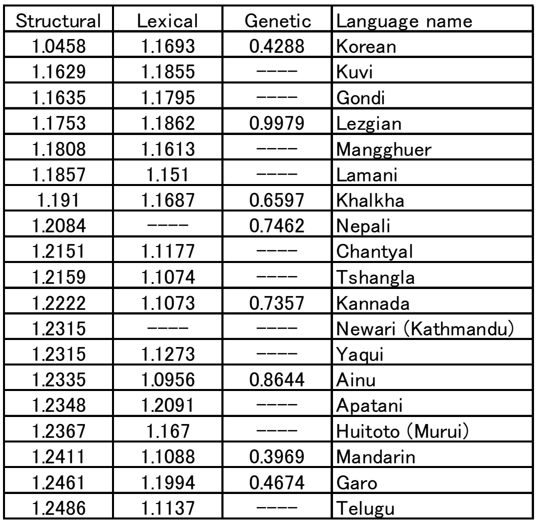
Comparison between three distances of languages are conducted using multi-dimensional scaling. They are structural, lexical and genetic distances. Correlation coefficients between any two of the three distance data are almost 0.2 that have weak correlations. Restricting data around Japanese, correlation coefficients are changed to low in structural and lexical distances and lexical and genetic distances. On the other hand, correlation coefficient for structural and genetic distances against Japanese is also almost 0.2.

Some structurally close languages against Japanese are overlapped to genetically close languages. They are Mandarin, Korean and Garo.

There are lots of future works. In this study, calculation methods of dissimilarities are simple. More sophisticated calculation methods can be considered. For the structural dissimilarity, all features are treated equally. WALS features are classified into eleven groups. Different dissimilarity metrics may be considerable for different feature groups. For the lexical dissimilarity, simple edit distance is used. More complex cost function may be usable. For example, similarity condition defined by Brown et al. (2008) can be considered. For genetic dissimilarity, granularity of YHg should be examined. Finer granularity and/or granularity suitable for distinguishing populations can be considered. Genetic distance metrics should be re-examined as the calculation method of genetic dissimilarity for the analysis.

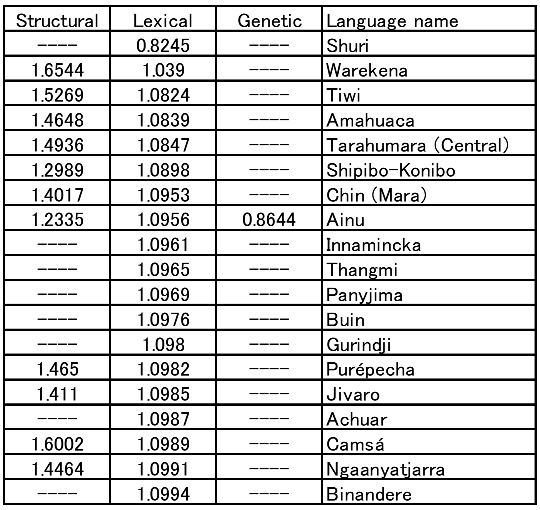
YHg reflects ancestors in the male line. On the other hand, female line is reflected to mitochondrial DNA (mtDNA). Data of mtDNA

Table 4: Top 20 languages close to Japanese (---- means no data)



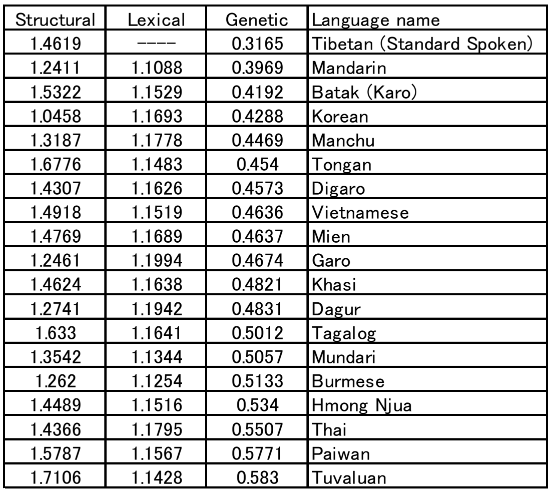
Structural	Lexical	Genetic	Language name
1.0458	1.1693	0.4288	Korean
1.1629	1.1855	----	Kuvi
1.1635	1.1795	----	Gondi
1.1753	1.1862	0.9979	Lezgian
1.1808	1.1613	----	Mangghuer
1.1857	1.151	----	Lamani
1.191	1.1687	0.6597	Khalkha
1.2084	----	0.7462	Nepali
1.2151	1.1177	----	Chantyal
1.2159	1.1074	----	Tshangla
1.2222	1.1073	0.7357	Kannada
1.2315	----	----	Newari (Kathmandu)
1.2315	1.1273	----	Yaqui
1.2335	1.0956	0.8644	Ainu
1.2348	1.2091	----	Apatani
1.2367	1.167	----	Huitoto (Murui)
1.2411	1.1088	0.3969	Mandarin
1.2461	1.1994	0.4674	Garo
1.2486	1.1137	----	Telugu

(a) Structural distance



Structural	Lexical	Genetic	Language name
----	0.8245	----	Shuri
1.6544	1.039	----	Warekena
1.5269	1.0824	----	Tiwi
1.4648	1.0839	----	Amahuaca
1.4936	1.0847	----	Tarahumara (Central)
1.2989	1.0898	----	Shipibo-Konibo
1.4017	1.0953	----	Chin (Mara)
1.2335	1.0956	0.8644	Ainu
----	1.0961	----	Innamincka
----	1.0965	----	Thangmi
----	1.0969	----	Panyjima
----	1.0976	----	Buin
----	1.098	----	Gurindji
1.465	1.0982	----	Purépecha
1.411	1.0985	----	Jivaro
----	1.0987	----	Achuar
1.6002	1.0989	----	Camsá
1.4464	1.0991	----	Ngaanyatjarra
----	1.0994	----	Binandere

(b) Lexical distance



Structural	Lexical	Genetic	Language name
1.4619	----	0.3165	Tibetan (Standard Spoken)
1.2411	1.1088	0.3969	Mandarin
1.5322	1.1529	0.4192	Batak (Karo)
1.0458	1.1693	0.4288	Korean
1.3187	1.1778	0.4469	Manchu
1.6776	1.1483	0.454	Tongan
1.4307	1.1626	0.4573	Digaro
1.4918	1.1519	0.4636	Vietnamese
1.4769	1.1689	0.4637	Mien
1.2461	1.1994	0.4674	Garo
1.4624	1.1638	0.4821	Khasi
1.2741	1.1942	0.4831	Dagur
1.633	1.1641	0.5012	Tagalog
1.3542	1.1344	0.5057	Mundari
1.262	1.1254	0.5133	Burmese
1.4489	1.1516	0.534	Hmong Njua
1.4366	1.1795	0.5507	Thai
1.5787	1.1567	0.5771	Paiwan
1.7106	1.1428	0.583	Tuvaluan

(c) Genetic distance

haplogroups by populations should be added in the analysis.

References

Cecil H. Brown, Eric W. Holman, Søren Wichmann and Viveka Velupillai. 2008. Automated Classification of the World's Languages: a Description of the Method and Preliminary Results, *STUF - Language Typology and Universals,* Vol.61, No.4, pages 285-308.

L. L. Cavalli-Sforza, Eric Minch and J.L. Mountain. 1992. Coevolution of Genes and Languages revisited, *Proceedings of National Academy of Science*, Vol.89, pages 5620-5624.

Matthew S. Dryer and Martin Haspelmath (eds.) 2013. The World Atlas of Language Structures Online. *Leipzig: Max Planck Institute for Evolutionary Anthropology.* (http://wals.info, Accessed on 2015-11-23).

Terumasa Ehara. 2009. Analysis of Languages around Japan using The World Atlas of Language Structures' data, *Proceedings of The 19th Annual Meeting of The Association for Natural Language Processing*, C3-6, pages 438-441. (In Japanese).

Eric W. Holman. 2014. Programs for calculating ASJP distance matrices (version 2.2), (http://asjp.clld.org/static/ASJPSoftware003.zip, Accessed on 2015-10-18).

Tatiana M. Karafet, Fernando L. Mendez, Monica B. Meilerman, Peter A. Underhill, Stephen L. Zegura, and Michael F. Hammer. 2008. New Binary Polymorphisms Reshape and Increase Resolution of the Human Y Chromosomal Haplogroup Tree, *Genome Research*, Vol. 18, pages 830-838.

Masatoshi Nei and A. K. Roychoudhury. 1974. Genic Variation Within and Between the Three Major Races of Man, Caucasoids, Negroids, and Mongoloids, *The American Journal of Human Genetics*, Vol.26, No.4, pages 421–443.

Maurizio Serva and Filippo Petroni. 2009. Indo-European languages tree by Levenshtein distance, *Europhysics Letters,* Vol. 81, No.6.

Søren Wichmann, André Müller, Annkathrin Wett, Viveka Velupillai, Julia Bischoffberger, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Zarina Molochieva, Pamela Brown, Harald Hammarström, Oleg Belyaev, Johann-Mattis List, Dik Bakker, Dmitry Egorov, Matthias Urban, Robert Mailhammer, Agustina Carrizo, Matthew S. Dryer, Evgenia Korovina, David Beck, Helen Geyer, Patience Epps, Anthony Grant, and Pilar Valenzuela. 2013. The ASJP Database (version 16). (http://asjp.clld.org/, Accessed on 2015-10-18).

Wikimedia Foundation. 2015. Y-chromosome haplogroups by populations, *Wikipedia.* (https://en.wikipedia.org/wiki/Y-chromosome_haplogroups_by_populations, Accessed on 2015-11-20).