Northern Language Studies, 14: 125-154 Supplementary Materials

Areality and Genealogy in Linguistic Typology:

A Graphical Modeling Approach to Pacific Rim

Yohei ONO
(St. Luke's International University)

Supplementary Materials

1. Why we need to transform original data into chi statistic

This Section will explain why we need to transform our original data into chi statistic from the viewpoints of linguistic typology. Since WALS contains numerous "missing values" that cannot be imputed by present value on corresponding feature, there are some biases in each language how many values the language contains and also there are some biases in each value in how many languages the value is included, which will lead us to revising "intuitive" measure between languages or values. In this Section, our objective is to understand the nature of biases in our original data and why chi statistic can revise these biases statistically, taking tentative data as examples.

Table 1 is a tentative example of data, where row names correspond to language name and column names correspond to value name, respectively. In Table 1, one intuitive similarity measure can be as follows: we count agreement of the corresponding values as 1 or disagreement of the corresponding values as 0 in pairs of languages or values. However, the intuitive similarity measure needs to be revised regarding some biases in the similarity data, which corresponds to the first requirement, "correction on frequencies" in our paper.

"Correction on frequencies" is originated from following biases. Since the numbers of values are not equal but biased in each language, some languages containing relatively larger numbers of values tend to be more similar to other languages and those languages are likely to be more similar to each other. Also, since the numbers of languages are not equal but biased in each value, some values included in relatively larger numbers of languages tend to be more similar to other values and those values are likely to be more similar to each other. Furthermore, some languages containing more "popular" values tend to be more similar to other languages and those languages are likely to be more similar to each other.

For example, Table 1 shows tentative example of data between languages and values, Table 2 shows calculated similarity between languages, and Table 3 shows calculated similarity between values. In Table 1, A language contains the largest number of values in all languages (i.e., 8 values), which resulting in A language more similar to other languages in Table 2. Furthermore, I language contains smaller numbers of values (i.e., 4 values) but they are more "popular" value (i.e. V1, V4, V7), which resulting in I language more similar to other languages.

Furthermore, Table 1 illustrated that V1 value is included in the largest number of languages (i.e., 8 languages), which resulting in V1 value more similar to other values in Table 3. Furthermore,

V5 value is included in smaller numbers of languages (i.e., 3 languages) but they contain relatively larger number of values (i.e. A, E, and G languages), which resulting in V5 value more similar to other values.

Table 1. Example of data. Row names correspond to language name, respectively. Colum names correspond to value name, respectively. Presence of value is coded as 1 and absence as 0. SUM corresponds to the total number of present values in each row or column, respectively.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	SUM
A	1	0	1	1	1	1	1	1	0	1	8
B	0	1	0	1	0	0	1	0	0	1	4
C	1	0	0	1	0	1	0	0	0	1	4
D	0	0	1	0	0	0	1	0	0	0	2
E	1	0	1	0	1	0	1	0	1	0	5
F	1	1	0	0	0	0	1	0	0	0	3
G	1	0	0	1	1	0	0	0	1	0	4
H	1	1	0	0	0	0	0	1	0	0	3
I	1	0	0	1	0	0	1	1	0	0	4
J	1	0	0	0	0	0	0	1	1	0	3
SUM	8	3	3	5	3	2	6	4	3	3	40

Table 2. Similarity between languages obtained from Table 1.

	A	B	C	D	E	F	G	H	I	J
A	8	3	4	2	4	2	3	2	4	2
B	3	4	2	1	1	2	1	1	2	0
C	4	2	4	0	1	1	2	1	2	1
D	2	1	0	2	2	1	0	0	1	0
E	4	1	1	2	5	2	3	1	2	2
F	2	2	1	1	2	3	1	2	2	1
G	3	1	2	0	3	1	4	1	2	2
H	2	1	1	0	1	2	1	3	2	2
I	4	2	2	1	2	2	2	2	4	2
J	2	0	1	0	2	1	2	2	2	3

Table 3. Similarity between values obtained from Table 1.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	8	2	2	4	3	2	4	4	3	2
V2	2	3	0	1	0	0	2	1	0	1
V3	2	0	3	1	2	1	3	1	1	1
V4	4	1	1	5	2	2	3	2	1	3
V5	3	0	2	2	3	1	2	1	2	1
V6	2	0	1	2	1	2	1	1	0	2
V7	4	2	3	3	2	1	6	2	1	2
V8	4	1	1	2	1	1	2	4	1	1
V9	3	0	1	1	2	0	1	1	3	0
V10	2	1	1	3	1	2	2	1	0	3

Thus, we need to revise our proposed data for removing these biases in order to capture the similarity between languages or values more accurately. As explained below, chi statistic is one promising solution to “correction on frequencies.”

Let us introduce some mathematical notations in Table 1.

N_{ij} is the number in i-th row and j-th column;
$N_{i\cdot}$ is the sum of the number in i-th row, and $N_{i\cdot} = \sum_{j=1}^{10} N_{ij}$ in our case;
$N_{\cdot j}$ is the sum of the number in j-th column, and $N_{\cdot j} = \sum_{i=1}^{10} N_{ij}$ in our case;
N is the total sum of number, and $N = \sum_{i=1}^{10} N_{i\cdot} = \sum_{j=1}^{10} N_{\cdot j}$ in our case;
P_{ij} is the relative frequency in i-th row and j-column, and $P_{ij} = N_{ij}/N$;
$P_{i\cdot}$ is the relative frequency in i-th row, and $P_{i\cdot} = N_{i\cdot}/N$;
$P_{\cdot j}$ is the relative frequency in j-th column, and $P_{\cdot j} = N_{\cdot j}/N$;
$i = 1, 2, \dots, 10$ and $j = 1, 2, \dots, 10$ in our case.

For example, $N_{11} = 1$, $N_{14} = 1$, $N_{41} = 1, N_{2\cdot} = 4, N_{\cdot 2} = 3, N = 40, P_{11} = N_{11}/N = 1/40 = 0.025, P_{14} = N_{14}/N = 1/40 = 0.025, P_{41} = N_{41}/N = 0/40 = 0, P_{2\cdot} = N_{2\cdot}/N = 4/40 = 0.1,$ $and\ P_{\cdot 2} = N_{\cdot 2}/N = 3/40 = 0.075$ in Table 1.

Chi statistic is defined as $\chi_{ij} = \left(N_{ij} - N \cdot P_{i\cdot} \cdot P_{\cdot j}\right)/\sqrt{N_{i\cdot} \cdot N_{\cdot j}}$ in i-th row and j-th column of Table 1, corresponding to each cell in Table 7.

We will try to understand why we need to transform our original data into chi statistic in the context of linguistic typology in following explanations. First, let us consider the meaning of numerator in chi statistic, $\left(N_{ij} - N \cdot P_{i\cdot} \cdot P_{\cdot j}\right)$.

For example, $N_{11} = 1$ (i.e., the number in the first row and the first column) and $N_{43} = 1$ (i.e., the number in the fourth row and the third column) are the same value but biased in Table 1 because A language contains the largest number of values (i.e., 8 values) in all languages and V1 value is included in the largest number of languages (i.e., 8 languages) in the case of $N_{11} = 1$ but D language contains the smallest number of values in all languages (i.e., 2 values) and V3 value is included in relatively smaller number of languages (i.e., 3 languages) in the case of $N_{43} = 1$. In other words, the value of 1 can occur much more likely in N_{11} than in N_{43}. Thus, the main objective of numerator in chi statistic is to revise the bias of occurring the data in Table 1.

In Table 4, we transformed Table 1 into the numerators in chi statistic, $\left(N_{ij} - N \cdot P_{i\cdot} \cdot P_{\cdot j}\right)$. For example, the numerator of χ_{11} (i.e., chi statistic in the first row and the first column) is calculated as $N_{11} - N \cdot P_{1\cdot} \cdot P_{\cdot 1} = 1 - 40 \cdot (N_{1\cdot}/N) \cdot (N_{\cdot 1}/N) = 1 - 40 \cdot (8/40) \cdot (8/40) = 1 - 1.6 = -0.6$, and the numerator of χ_{43} (i.e., chi statistic in in the fourth row and the third column) is calculated as $N_{43} - N \cdot P_{4\cdot} \cdot P_{\cdot 3} = 1 - 40 \cdot (N_{4\cdot}/N) \cdot (N_{\cdot 3}/N) = 1 - 40 \cdot (2/40) \cdot (3/40) = 1 - 0.15 = 0.85$.

Notably, the correction term is defined as $N \cdot P_{i\cdot} \cdot P_{\cdot j}$ in the numerator of chi statistic, corresponding to the "expected" or "average" number in i-th row and j-th column. Thus, the numerator of chi statistic (i.e., $N_{ij} - N \cdot P_{i\cdot} \cdot P_{\cdot j}$) can revise the bias in the data by subtracting "expected" or "average" number from original one in each combination of row and column, where the "expected" or "average" number will reflect the biases by heterogeneous total numbers in row and column explained above. For example, the numerator of χ_{11} is negative value (i.e., -0.6) and the numerator of χ_{43} is positive value (i.e., 0.85).

The former suggested the similarity between the first row and the first column is weak by negative value because $N_{11} = 1$ can be strongly affected by the total number of the first row and the first column (i.e., $N_{1\cdot} = 8$ and $N_{\cdot 1} = 8$). Furthermore, the latter suggested the similarity between the

fourth row and the third column is strong by positive value because $N_{43} = 1$ can be weakly affected by the total number of the first row and the first column (i.e., $N_{4\cdot} = 2$ and $N_{\cdot 3} = 3$).

Thus, the numerator in chi statistic, $\left(N_{ij} - N \cdot P_{i\cdot} \cdot P_{\cdot j}\right)$ can be considered as satisfying "correction on frequencies" in our paper. However, further revisions are needed in the numerator in chi statistic as explained in Table 5.

Table 4. The numerators on chi statistic from Table 1.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A	-0.6	-0.6	0.4	0	0.4	0.6	-0.2	0.2	-0.6	0.4
B	-0.8	0.7	-0.3	0.5	-0.3	-0.2	0.4	-0.4	-0.3	0.7
C	0.2	-0.3	-0.3	0.5	-0.3	0.8	-0.6	-0.4	-0.3	0.7
D	-0.4	-0.15	0.85	-0.25	-0.15	-0.1	0.7	-0.2	-0.15	-0.15
E	0	-0.38	0.625	-0.63	0.625	-0.25	0.25	-0.5	0.625	-0.38
F	0.4	0.775	-0.23	-0.38	-0.23	-0.15	0.55	-0.3	-0.23	-0.23
G	0.2	-0.3	-0.3	0.5	0.7	-0.2	-0.6	-0.4	0.7	-0.3
H	0.4	0.775	-0.23	-0.38	-0.23	-0.15	-0.45	0.7	-0.23	-0.23
I	0.2	-0.3	-0.3	0.5	-0.3	-0.2	0.4	0.6	-0.3	-0.3
J	0.4	-0.23	-0.23	-0.38	-0.23	-0.15	-0.45	0.7	0.775	-0.23

Let us introduce some mathematical notations in Table 5.

N'_{ij} is the number in i-th row and j-th column;

$N'_{i\cdot}$ is the sum of the number in i-th row, and $N'_{i\cdot} = \sum_{j=1}^{10} N'_{ij}$ in our case;

$N'_{\cdot j}$ is the sum of the number in j-th column, and $N'_{\cdot j} = \sum_{i=1}^{10} N'_{ij}$ in our case;

N' is the total sum of number, and $N' = \sum_{i=1}^{10} N'_{i\cdot} = \sum_{j=1}^{10} N'_{\cdot j}$ in our case;

P'_{ij} is the relative frequency in i-th row and j-column, and $P'_{ij} = N'_{ij}/N'$;

$P'_{i\cdot}$ is the relative frequency in i-th row, and $P'_{i\cdot} = N'_{i\cdot}/N'$;

$P'_{\cdot j}$ is the relative frequency in j-th column, and $P'_{\cdot j} = N'_{\cdot j}/N'$;

$i = 1, 2, \ldots, 10$ and $j = 1, 2, \ldots, 10$ in our case.

For example, $N'_{11} = 1$, $N'_{14} = 0$, $N'_{41} = 0, N'_{2\cdot} = 2, N'_{\cdot 2} = 1, N' = 20, P'_{11} = N'_{11}/N' = 1/20 = 0.05, P'_{14} = N'_{14}/N' = 0/20 = 0, P'_{41} = N'_{41}/N' = 0/20 = 0, P'_{2\cdot} = N'_{2\cdot}/N' = 2/20 = 0.1,$ $and\ P'_{\cdot 2} = N'_{\cdot 2}/N' = 1/20 = 0.05$ in Table 5. Also, Table 6 is the numerator in chi statistic calculated from Table 6 (i.e., $\chi'_{ij} = N'_{ij} - N' \cdot P'_{i\cdot} \cdot P'_{\cdot j}$).

Let us focus on the numbers in the first row and the third column in Table 4 and Table 6, respectively. The numerator in chi statistic is calculated in Table 4 as $\chi_{13} = N_{13} - N \cdot P_{1\cdot} \cdot P_{\cdot 3} = 1 - 40 \cdot (N_{1\cdot}/N) \cdot (N_{\cdot 3}/N) = 1 - 40 \cdot (8/40) \cdot (3/40) = 1 - 0.6 = 0.4$, and in Table 5 as $\chi'_{13} = N'_{13} - N' \cdot P'_{1\cdot} \cdot P'_{\cdot 3} = 1 - 20 \cdot (N'_{1\cdot}/N') \cdot (N'_{\cdot 3}/N') = 1 - 20 \cdot (4/20) \cdot (3/20) = 1 - 0.6 = 0.4$. Thus, the numerators in chi statistic are the same on the first row and the third column in Table 4 and Table 6, suggesting that these revised similarities are the same degree of association between the first row and the third column in Table 4 and Table 6.

However, $P'_{1\cdot} = N'_{1\cdot}/N' = 4/20 = 0.2$ and $P'_{\cdot 3} = N'_{\cdot 3}/N' = 3/20 = 0.15$ in Table 6, whereas $P_{1\cdot} = N_{1\cdot}/N = 8/40 = 0.2$ and $P_{\cdot 3} = N_{\cdot 3}/N = 3/40 = 0.075$ in Table 4. Thus, the numerator in chi statistic is more affected by the total number of the third column in Table 6 than in Table 4, which will require us further reconsidering the numerator in chi statistic.

Table 5. Example of data. Row names correspond to language name, respectively. Colum names correspond to value name, respectively. Presence of value is coded as 1 and absence as 0. SUM corresponds to the total number of present values in each row or column, respectively.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	SUM
A	1	0	1	0	0	0	1	1	0	0	4
B	0	1	0	0	0	0	0	0	0	1	2
C	1	0	0	0	0	0	0	0	0	1	2
D	0	0	1	0	0	0	0	0	0	0	1
E	0	0	1	0	0	0	0	0	1	0	2
F	1	0	0	0	1	0	1	0	0	0	3
G	0	0	0	1	0	0	0	0	0	0	1
H	0	0	0	0	0	1	0	0	0	0	1
I	0	0	0	1	0	0	0	1	0	0	2
J	1	0	0	0	0	0	0	1	0	0	2
SUM	4	1	3	2	1	1	2	3	1	2	20

Table 6. The numerators on chi statistic from Table 5.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A	0.2	-0.2	0.4	-0.4	-0.2	-0.2	0.6	0.4	-0.2	-0.4
B	-0.4	0.9	-0.3	-0.2	-0.1	-0.1	-0.2	-0.3	-0.1	0.8
C	0.6	-0.1	-0.3	-0.2	-0.1	-0.1	-0.2	-0.3	-0.1	0.8
D	-0.2	-0.05	0.85	-0.1	-0.05	-0.05	-0.1	-0.15	-0.05	-0.1
E	-0.4	-0.1	0.7	-0.2	-0.1	-0.1	-0.2	-0.3	0.9	-0.2
F	0.4	-0.15	-0.45	-0.3	0.85	-0.15	0.7	-0.45	-0.15	-0.3
G	-0.2	-0.05	-0.15	0.9	-0.05	-0.05	-0.1	-0.15	-0.05	-0.1
H	-0.6	-0.15	-0.45	-0.3	-0.15	0.85	-0.3	-0.45	-0.15	-0.3
I	-0.4	-0.1	-0.3	0.8	-0.1	-0.1	-0.2	0.7	-0.1	-0.2
J	0.6	-0.1	-0.3	-0.2	-0.1	-0.1	-0.2	0.7	-0.1	-0.2

Chi statistic will enable us to correct the biases in the numerator in chi statistic, dividing the numerator of chi statistic by square root of product for the total number of the corresponding row and column. Thus, chi statistic is defined in Table 1 as $\chi_{ij} = \left(N_{ij} - N \cdot P_{i\cdot} \cdot P_{\cdot j}\right) / \sqrt{N_{i\cdot} \cdot N_{\cdot j}}$ and in Table 5 as $\chi'_{ij} = \left(N'_{ij} - N' \cdot P'_{i\cdot} \cdot P'_{\cdot j}\right) / \sqrt{N'_{i\cdot} \cdot N'_{\cdot j}}$

Table 7. Chi statistics obtained from Table 1.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A	-0.075	-0.122	0.082	0	0.082	0.150	-0.029	0.035	-0.122	0.082
B	-0.141	0.202	-0.087	0.112	-0.087	-0.071	0.082	-0.100	-0.087	0.202
C	0.035	-0.087	-0.087	0.112	-0.087	0.283	-0.122	-0.100	-0.087	0.202
D	-0.100	-0.061	0.347	-0.079	-0.061	-0.050	0.202	-0.071	-0.061	-0.061
E	0	-0.097	0.161	-0.125	0.161	-0.079	0.046	-0.112	0.161	-0.097
F	0.082	0.258	-0.075	-0.097	-0.075	-0.061	0.130	-0.087	-0.075	-0.075
G	0.035	-0.087	-0.087	0.112	0.202	-0.071	-0.122	-0.100	0.202	-0.087
H	0.082	0.258	-0.075	-0.097	-0.075	-0.061	-0.106	0.202	-0.075	-0.075
I	0.035	-0.087	-0.087	0.112	-0.087	-0.071	0.082	0.150	-0.087	-0.087
J	0.082	-0.075	-0.075	-0.097	-0.075	-0.061	-0.106	0.202	0.258	-0.075

Table 8. Chi statistics obtained from Table 5.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A	0.050	-0.100	0.115	-0.141	-0.100	-0.100	0.212	0.115	-0.100	-0.141
B	-0.141	0.636	-0.122	-0.100	-0.071	-0.071	-0.100	-0.122	-0.071	0.400
C	0.212	-0.071	-0.122	-0.100	-0.071	-0.071	-0.100	-0.122	-0.071	0.400
D	-0.100	-0.050	0.491	-0.071	-0.050	-0.050	-0.071	-0.087	-0.050	-0.071
E	-0.141	-0.071	0.286	-0.100	-0.071	-0.071	-0.100	-0.122	0.636	-0.100
F	0.115	-0.087	-0.150	-0.122	0.491	-0.087	0.286	-0.150	-0.087	-0.122
G	-0.100	-0.050	-0.087	0.636	-0.050	-0.050	-0.071	-0.087	-0.050	-0.071
H	-0.100	-0.050	-0.087	-0.071	-0.050	0.950	-0.071	-0.087	-0.050	-0.071
I	-0.141	-0.071	-0.122	0.400	-0.071	-0.071	-0.100	0.286	-0.071	-0.100
J	0.212	-0.071	-0.122	-0.100	-0.071	-0.071	-0.100	0.286	-0.071	-0.100

For example, chi statistic is calculated as $\chi_{13} = (N_{13} - N \cdot P_{1\cdot} \cdot P_{\cdot 3})/\sqrt{N_{1\cdot} \cdot N_{\cdot 3}} = (0.4)/\sqrt{8 \cdot 3} = 0.4/\sqrt{24} = 0.082$ in Table 7 and $\chi'_{13} = (N'_{13} - N' \cdot P'_{1\cdot} \cdot P'_{\cdot 3})/\sqrt{N'_{1\cdot} \cdot N'_{\cdot 3}} = (0.4)/\sqrt{4 \cdot 3} = 0.4/\sqrt{12} = 0.115$ in Table 8. In other words, chi statistic is larger in the first row and the third column on Table 8 than in Table 7, suggesting that the degree of association between the first row and the third column is stronger in Table 8 than in Table 7.

Thus, chi statistic is one useful measure satisfying "correction on frequencies" in this paper.

2. Why we need to develop chi statistics into multidimensional vectors

In previous section, we attempted to intuitively understand chi statistic in the context of "correction on frequencies" in this paper. As explained in Section 5.1, "correspondence analysis (Benzécri et coll. 1973) is a statistical tool that will decompose chi statistics into multidimensional vectors of both languages and values where each chi statistic is represented as inner product between the vector of the corresponding language and that of the corresponding value." The main objective of this section is to intuitively understand why we need to develop chi statistics into multidimensional vectors and how correspondence analysis will work in practice. Note that interested reader should refer to Kimiyama (2011) in Japanese literature.

In previous section, "correction on frequencies" required us transforming our original data (e.g., Table 1) into chi statistics (e.g., Table 7), whose numerator (i.e., $N_{ij} - N \cdot P_{i\cdot} \cdot P_{\cdot j}$) consists of the original number in i-th row and j-th column (i.e., N_{ij}) and the "expected" or "average" number in i-th row and j-th column (i.e., $N \cdot P_{i\cdot} \cdot P_{\cdot j}$). Notably, "correction on frequencies" resulted in subtracting the "expected" or "average" number from the original number, which led the original data to some degree of association between row and column by revising the effect of heterogeneous total numbers in row and column with the term of $-N \cdot P_{i\cdot} \cdot P_{\cdot j}$ and $\sqrt{N_{i\cdot} \cdot N_{\cdot j}}$ in chi statistic.

However, our main objective is to capture the characteristics of language or value from the chi statistics, and the characteristics of language or value is considered as "multidimensional" dispositions in linguistic typology. Since the degree of association between row and column corresponds to the meaning of the chi statistic, correspondence analysis attempts to assign numbers to each row and column, where the product of the assigned number is expected to approximate the corresponding chi statistic in each combination of row and column.

Table 9. Assigned numbers by correspondence analysis applied to Table 7.

		V1_1	V2_1	V3_1	V4_1	V5_1	V6_1	V7_1	V8_1	V9_1	V10_1
		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A1	A	-0.075	-0.122	0.082	0.000	0.082	0.150	-0.029	0.035	-0.122	0.082
B1	B	-0.141	0.202	-0.087	0.112	-0.087	-0.071	0.082	-0.100	-0.087	0.202
C1	C	0.035	-0.087	-0.087	0.112	-0.087	0.283	-0.122	-0.100	-0.087	0.202
D1	D	-0.100	-0.061	0.347	-0.079	-0.061	-0.050	0.202	-0.071	-0.061	-0.061
E1	E	0.000	-0.097	0.161	-0.125	0.161	-0.079	0.046	-0.112	0.161	-0.097
F1	F	0.082	0.258	-0.075	-0.097	-0.075	-0.061	0.130	-0.087	-0.075	-0.075
G1	G	0.035	-0.087	-0.087	0.112	0.202	-0.071	-0.122	-0.100	0.202	-0.087
H1	H	0.082	0.258	-0.075	-0.097	-0.075	-0.061	-0.106	0.202	-0.075	-0.075
I1	I	0.035	-0.087	-0.087	0.112	-0.087	-0.071	0.082	0.150	-0.087	-0.087
J1	J	0.082	-0.075	-0.075	-0.097	-0.075	-0.061	-0.106	0.202	0.258	-0.075

Table 10. Products of assigned numbers by correspondence analysis applied to Table 7.

		V1_1	V2_1	V3_1	V4_1	V5_1	V6_1	V7_1	V8_1	V9_1	V10_1
		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A1	A	L1*A1*V1_1	L1*A1*V2_1	L1*A1*V3_1	L1*A1*V4_1	L1*A1*V5_1	L1*A1*V6_1	L1*A1*V7_1	L1*A1*V8_1	L1*A1*V9_1	L1*A1*V10_1
B1	B	L1*B1*V1_1	L1*B1*V2_1	L1*B1*V3_1	L1*B1*V4_1	L1*B1*V5_1	L1*B1*V6_1	L1*B1*V7_1	L1*B1*V8_1	L1*B1*V9_1	L1*B1*V10_1
C1	C	L1*C1*V1_1	L1*C1*V2_1	L1*C1*V3_1	L1*C1*V4_1	L1*C1*V5_1	L1*C1*V6_1	L1*C1*V7_1	L1*C1*V8_1	L1*C1*V9_1	L1*C1*V10_1
D1	D	L1*D1*V1_1	L1*D1*V2_1	L1*D1*V3_1	L1*D1*V4_1	L1*D1*V5_1	L1*D1*V6_1	L1*D1*V7_1	L1*D1*V8_1	L1*D1*V9_1	L1*D1*V10_1
E1	E	L1*E1*V1_1	L1*E1*V2_1	L1*E1*V3_1	L1*E1*V4_1	L1*E1*V5_1	L1*E1*V6_1	L1*E1*V7_1	L1*E1*V8_1	L1*E1*V9_1	L1*E1*V10_1
F1	F	L1*F1*V1_1	L1*F1*V2_1	L1*F1*V3_1	L1*F1*V4_1	L1*F1*V5_1	L1*F1*V6_1	L1*F1*V7_1	L1*F1*V8_1	L1*F1*V9_1	L1*F1*V10_1
G1	G	L1*G1*V1_1	L1*G1*V2_1	L1*G1*V3_1	L1*G1*V4_1	L1*G1*V5_1	L1*G1*V6_1	L1*G1*V7_1	L1*G1*V8_1	L1*G1*V9_1	L1*G1*V10_1
H1	H	L1*H1*V1_1	L1*H1*V2_1	L1*H1*V3_1	L1*H1*V4_1	L1*H1*V5_1	L1*H1*V6_1	L1*H1*V7_1	L1*H1*V8_1	L1*H1*V9_1	L1*H1*V10_1
I1	I	L1*I1*V1_1	L1*I1*V2_1	L1*I1*V3_1	L1*I1*V4_1	L1*I1*V5_1	L1*I1*V6_1	L1*I1*V7_1	L1*I1*V8_1	L1*I1*V9_1	L1*I1*V10_1
J1	J	L1*J1*V1_1	L1*J1*V2_1	L1*J1*V3_1	L1*J1*V4_1	L1*J1*V5_1	L1*J1*V6_1	L1*J1*V7_1	L1*J1*V8_1	L1*J1*V9_1	L1*J1*V10_1

Table 9 is an intuitive image on assigned numbers to each row and column in Table 7, where the number of A1 is assigned to A, the number of B1 is assigned to B and the same analysis can apply to D1-J1, and the number of V1_1 is assigned to V1, the number of V2_1 is assigned to V2, and the same analysis can apply to V3_1-V10_1. In practice, correspondence analysis will also calculate L1 and approximate Table 7 as shown in Table 10. For example, χ_{11} (i.e., the number in the first row and the first column) is approximated by L1 * A1 * V1_1, χ_{12} (i.e., the number in the first row and the second column) is approximated by L1 * A1 * V2_1, and the same analysis can apply to each combination of row and column.

Correspondence analysis will attempt to calculate these quantities (i.e., A1-J1, V1_1-V10_1, and L1) in order to minimize sum of differences between the numbers of Table 7 and Table 10, applying the algorithm of singular value decomposition (abbreviated as SVD hereafter) to Table 7.

Table 11 is the results on assigned numbers and their products by SVD applied to Table 7. In this table, A1 corresponds to -0.022, B1 to -0.474, and the same analysis can apply to other numbers. Also, V1_1 corresponds to 0.048, V2_1 to -0.478, and the same analysis can apply to other numbers. Furthermore, L1 is calculated as 0.628 in Table 11.

Thus, Table 11 illustrated that the number in Table 11 is not the same in Table 7, which demonstrated that chi statistics in Table 7 do not capture in either languages or values by just one series of numbers. Note that a series of numbers is called sometimes as “vector” in mathematics.

Table 11. Calculated numbers and their product by correspondence analysis applied to Table 7.

Note that these numbers are rounded in fourth decimal in following tables.

L1 = 0.628		0.048	-0.478	0.341	-0.193	0.372	-0.204	-0.029	-0.008	0.536	-0.384
		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
-0.022	A	-0.001	0.007	-0.005	0.003	-0.005	0.003	0.000	0.000	-0.007	0.005
-0.474	B	-0.014	0.142	-0.101	0.057	-0.111	0.061	0.008	0.002	-0.160	0.114
-0.346	C	-0.010	0.104	-0.074	0.042	-0.081	0.044	0.006	0.002	-0.117	0.084
0.208	D	0.006	-0.063	0.045	-0.025	0.049	-0.027	-0.004	-0.001	0.070	-0.050
0.517	E	0.016	-0.155	0.111	-0.063	0.121	-0.066	-0.009	-0.003	0.174	-0.125
-0.249	F	-0.008	0.075	-0.053	0.030	-0.058	0.032	0.004	0.001	-0.084	0.060
0.362	G	0.011	-0.109	0.077	-0.044	0.085	-0.046	-0.006	-0.002	0.122	-0.087
-0.242	H	-0.007	0.073	-0.052	0.029	-0.056	0.031	0.004	0.001	-0.081	0.058
-0.067	I	-0.002	0.020	-0.014	0.008	-0.016	0.009	0.001	0.000	-0.023	0.016
0.297	J	0.009	-0.089	0.063	-0.036	0.069	-0.038	-0.005	-0.001	0.100	-0.072

In other words, we calculated two vectors in this procedure; one corresponds to the languages (i.e., the assigned numbers on row), the other to the values (i.e., the assigned numbers on column).

Since our calculated vectors described only part of chi statistic in Table 7, we will apply singular value decomposition to the "residuals", that is, the data obtained by subtracting Table 11 from Table 7 in each combination of row and column as shown in Table 12.

Table 12. Assigned numbers by correspondence analysis applied to the "residuals" between Table 7 and Table 11.

		V1_2	V2_2	V3_2	V4_2	V5_2	V6_2	V7_2	V8_2	V9_2	V10_2
		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A2	A	-0.074	-0.129	0.086	-0.003	0.087	0.147	-0.029	0.035	-0.115	0.076
B2	B	-0.127	0.060	0.015	0.054	0.024	-0.131	0.073	-0.102	0.073	0.088
C2	C	0.046	-0.191	-0.012	0.070	-0.006	0.239	-0.129	-0.102	0.030	0.118
D2	D	-0.106	0.001	0.302	-0.054	-0.110	-0.023	0.206	-0.070	-0.131	-0.011
E2	E	-0.016	0.058	0.051	-0.062	0.041	-0.013	0.055	-0.109	-0.013	0.028
F2	F	0.089	0.184	-0.022	-0.127	-0.017	-0.093	0.125	-0.088	0.009	-0.135
G2	G	0.024	0.022	-0.164	0.156	0.118	-0.024	-0.116	-0.098	0.080	0.001
H2	H	0.089	0.186	-0.023	-0.126	-0.019	-0.092	-0.110	0.201	0.006	-0.133
I2	I	0.037	-0.107	-0.072	0.104	-0.071	-0.079	0.080	0.150	-0.064	-0.103
J2	J	0.073	0.014	-0.138	-0.061	-0.144	-0.023	-0.101	0.204	0.158	-0.003

Table 13. Products of assigned numbers by correspondence analysis applied to Table 13.

		V1_2	V2_2	V3_2	V4_2	V5_2	V6_2	V7_2	V8_2	V9_2	V10_2
		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
A2	A	L2*A2*V1_2	L2*A2*V2_2	L2*A2*V3_2	L2*A2*V4_2	L2*A2*V5_2	L2*A2*V6_2	L2*A2*V7_2	L2*A2*V8_2	L2*A2*V9_2	L2*A2*V10_2
B2	B	L2*B2*V1_2	L2*B2*V2_2	L2*B2*V3_2	L2*B2*V4_2	L2*B2*V5_2	L2*B2*V6_2	L2*B2*V7_2	L2*B2*V8_2	L2*B2*V9_2	L2*B2*V10_2
C2	C	L2*C2*V1_2	L2*C2*V2_2	L2*C2*V3_2	L2*C2*V4_2	L2*C2*V5_2	L2*C2*V6_2	L2*C2*V7_2	L2*C2*V8_2	L2*C2*V9_2	L2*C2*V10_2
D2	D	L2*D2*V1_2	L2*D2*V2_2	L2*D2*V3_2	L2*D2*V4_2	L2*D2*V5_2	L2*D2*V6_2	L2*D2*V7_2	L2*D2*V8_2	L2*D2*V9_2	L2*D2*V10_2
E2	E	L2*E2*V1_2	L2*E2*V2_2	L2*E2*V3_2	L2*E2*V4_2	L2*E2*V5_2	L2*E2*V6_2	L2*E2*V7_2	L2*E2*V8_2	L2*E2*V9_2	L2*E2*V10_2
F2	F	L2*F2*V1_2	L2*F2*V2_2	L2*F2*V3_2	L2*F2*V4_2	L2*F2*V5_2	L2*F2*V6_2	L2*F2*V7_2	L2*F2*V8_2	L2*F2*V9_2	L2*F2*V10_2
G2	G	L2*G2*V1_2	L2*G2*V2_2	L2*G2*V3_2	L2*G2*V4_2	L2*G2*V5_2	L2*G2*V6_2	L2*G2*V7_2	L2*G2*V8_2	L2*G2*V9_2	L2*G2*V10_2
H2	H	L2*H2*V1_2	L2*H2*V2_2	L2*H2*V3_2	L2*H2*V4_2	L2*H2*V5_2	L2*H2*V6_2	L2*H2*V7_2	L2*H2*V8_2	L2*H2*V9_2	L2*H2*V10_2
I2	I	L2*I2*V1_2	L2*I2*V2_2	L2*I2*V3_2	L2*I2*V4_2	L2*I2*V5_2	L2*I2*V6_2	L2*I2*V7_2	L2*I2*V8_2	L2*I2*V9_2	L2*I2*V10_2
J2	J	L2*J2*V1_2	L2*J2*V2_2	L2*J2*V3_2	L2*J2*V4_2	L2*J2*V5_2	L2*J2*V6_2	L2*J2*V7_2	L2*J2*V8_2	L2*J2*V9_2	L2*J2*V10_2

Table 13 is an intuitive image on assigned numbers to each row and column in Table 12, where the number of A2 is assigned to A, the number of B2 is assigned to B and the same analysis can apply to D2-J2, and the number of V1_2 is assigned to V1, the number of V2_2 is assigned to V2, and the same analysis can apply to V3_2-V10_2. Also, correspondence analysis will calculate L2 and

approximate Table 12 as shown in Table 13. For example, the residual is approximated by L2 * A2 * V1_2 in the first row and the first column, the residual is approximated by L2 * A2 * V2_2 in the first row and the second column, and the same analysis can apply to each combination of row and column.

Table 14. Calculated numbers and their product by correspondence analysis applied to Table 12.

L2 = 0.595		0.296	0.395	-0.432	-0.166	-0.122	-0.347	-0.189	0.445	0.254	-0.329
		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
-0.344	A	-0.061	-0.081	0.088	0.034	0.025	0.071	0.039	-0.091	-0.052	0.067
-0.095	B	-0.017	-0.022	0.024	0.009	0.007	0.020	0.011	-0.025	-0.014	0.019
-0.340	C	-0.060	-0.080	0.087	0.034	0.025	0.070	0.038	-0.090	-0.051	0.067
-0.388	D	-0.068	-0.091	0.100	0.038	0.028	0.080	0.044	-0.103	-0.059	0.076
-0.109	E	-0.019	-0.026	0.028	0.011	0.008	0.023	0.012	-0.029	-0.017	0.021
0.249	F	0.044	0.058	-0.064	-0.024	-0.018	-0.051	-0.028	0.066	0.038	-0.049
0.090	G	0.016	0.021	-0.023	-0.009	-0.007	-0.019	-0.010	0.024	0.014	-0.018
0.539	H	0.095	0.127	-0.139	-0.053	-0.039	-0.111	-0.061	0.143	0.081	-0.106
0.148	I	0.026	0.035	-0.038	-0.015	-0.011	-0.031	-0.017	0.039	0.022	-0.029
0.460	J	0.081	0.108	-0.118	-0.045	-0.033	-0.095	-0.052	0.122	0.069	-0.090

Table 14 is the results on assigned numbers and their products by SVD applied to Table 12. In this table, A2 corresponds to -0.344, B2 to -0.095, and the same analysis can apply to other numbers. Also, V1_2 corresponds to 0.296, V2_2 to -0.395, and the same analysis can apply to other numbers. Furthermore, L2 is calculated as 0.595 in Table 14.

Since our calculated vectors described only part of chi statistic in Table 12, we will repeat singular value decomposition to the “residuals”, that is, the data obtained by subtracting Table 14 from Table 14 in each combination of row and column.

Thus, these iterative procedures will calculate each vector on the corresponding language or value that can be considered as illustrating their multidimensional dispositions. For example, A language is described as the vector consisting of (-0.022, -0.344, …), in which the number in the first dimension is calculated as -0.022 in Table 11 and that in in the second dimension as -0.344 etc. Also, V1 value is described as the vector consisting of (0.048, 0.296, …), in which the number in the first dimension is calculated as 0.048 in Table 11 and that in in the second dimension as 0.296 etc. Therefore, the same analysis will work on each vector of language or value.

Thus, we can obtain the vectors as the multidimensional disposition of each language or value by applying correspondence analysis as explained above. In practice, the vectors are weighted by singular value in each dimension (e.g., L1 in Table 11 and L2 in Table 14) and some additional quantities. Note that interested reader can refer to Greenacre (2017) in this topic.

3. Why we cannot apply correlation coefficient to original binary data in WALS

In Section 5.2, we have introduced “profile view” and discussed that “profile view” is not based on Euclidean or Manhattan distance but based on correlation coefficient, which resulted in correlation coefficient playing significant role in our paper. However, some reader may pose a question why we cannot apply correlation coefficient to our original data, which is the main focus in this section.

Therefore, the main objective of this section is to illustrate intuitively why we cannot apply correlation coefficient to original data in WALS. Let us introduce some mathematical notations.

D_s: our data of D language in Table 1;

F_s: our data of F language in Table 1;

$\overline{D_s}$: mean of D_s, $\overline{D_s} = (\sum_{s=1}^{10} D_s)/10$;

$\overline{F_s}$: mean of F_s, $\overline{F_s} = (\sum_{s=1}^{10} F_s)/10$;

$s = 1, 2, \dots, 10$ in our case.

Thus, $D_s = (0,0,1,0,0,0,1,0,0,0)$, $F_s = (1,1,0,0,0,0,1,0,0,0)$, $\overline{D_s} = (\sum_{s=1}^{10} D_s)/10 = 2/10 = 0.2$, and $\overline{F_s} = (\sum_{s=1}^{10} F_s)/10 = 3/10 = 0.3$ in Table 1.

Correlation coefficient between D_s and F_s is defined as $r_{D_sF_s}$:

$$r_{D_sF_s} = \frac{\sum_{s=1}^{10}\{(D_s - \overline{D_s}) \cdot (F_s - \overline{F_s})\}}{\sqrt{\sum_{s=1}^{10}(D_s - \overline{D_s})^2 \cdot \sum_{s=1}^{10}(F_s - \overline{F_s})^2}}$$

In our case, $r_{D_sF_s}$ is calculated as follows:

$$\sum_{s=1}^{10}\{(D_s - \overline{D_s}) \cdot (F_s - \overline{F_s})\}$$
$$= \{(0-0.2)\cdot(1-0.3)\} + \{(0-0.2)\cdot(1-0.3)\} + \{(1-0.2)\cdot(0-0.3)\}$$
$$+\{(0-0.2)\cdot(0-0.3)\} + \{(0-0.2)\cdot(0-0.3)\} + \{(0-0.2)\cdot(0-0.3)\} + \{(1-0.2)\cdot(1-0.3)\}$$
$$+\{(0-0.2)\cdot(0-0.3)\} + \{(0-0.2)\cdot(0-0.3)\} + \{(0-0.2)\cdot(0-0.3)\}$$
$$= -0.14 + (-0.14) + (-0.24) + 0.06 + 0.06 + 0.06 + 0.56 + 0.06 + 0.06 + 0.06 = 0.4$$
$$\sum_{s=1}^{10}(D_s - \overline{D_s})^2$$
$$= (0-0.2)^2 + (0-0.2)^2 + (1-0.2)^2 + (0-0.2)^2 + (0-0.2)^2$$
$$+(0-0.2)^2 + (1-0.2)^2 + (0-0.2)^2 + (0-0.2)^2 + (0-0.2)^2 = 1.6$$
$$\sum_{s=1}^{10}(F_s - \overline{F_s})^2$$
$$= (1-0.3)^2 + (1-0.3)^2 + (0-0.3)^2 + (0-0.3)^2 + (0-0.3)^2$$
$$+(0-0.3)^2 + (1-0.3)^2 + (0-0.3)^2 + (0-0.3)^2 + (0-0.3)^2 = 2.1$$

$$r_{D_sF_s} = \frac{\sum_{s=1}^{10}\{(D_s - \overline{D_s}) \cdot (F_s - \overline{F_s})\}}{\sqrt{\sum_{s=1}^{10}(D_s - \overline{D_s})^2 \cdot \sum_{s=1}^{10}(F_s - \overline{F_s})^2}} = \frac{0.4}{\sqrt{1.6 \cdot 2.1}} = 0.218$$

Correlation coefficient aims to measure the degree of association between data (i.e., D language and F language in our case). However, the numerator of correlation coefficient (i.e., $\sum_{s=1}^{10}\{(D_s - \overline{D_s}) \cdot (F_s - \overline{F_s})\}$) is severely biased by the means of each data in the case that the data is binary or qualitative.

The bias in the numerator of correlation coefficient is illustrated in Figure 1, where the horizontal axis corresponds to data in D_s (i.e.., 0 or 1) and the vertical axis to data in F_s (i.e.., 0 or 1). For example, (D_s, F_s) is $(D_1, F_1) = (0,1)$ if $s = 1$, (D_s, F_s) is $(D_2, F_2) = (0,1)$ if $s = 2$, (D_s, F_s) is $(D_3, F_3) = (1,0)$ if $s = 3$ etc. Thus, the numerator of correlation coefficient (i.e., $\sum_{s=1}^{10}\{(D_s - \overline{D_s}) \cdot (F_s - \overline{F_s})\}$) is equal to the area of rectangular A if $(D_s, F_s) = (0,1)$, the area of rectangular B if $(D_s, F_s) = (1,1)$, the area of rectangular C if $(D_s, F_s) = (1,0)$, and the area of rectangular D if $(D_s, F_s) = (0,0)$. Thus, the sign of the numerator of correlation coefficient (i.e., $\sum_{s=1}^{10}\{(D_s - \overline{D_s}) \cdot (F_s - \overline{F_s})\}$) is plus if $(D_s, F_s) = (1,1)$ or $(D_s, F_s) = (0,0)$ (i.e., rectangular B or rectangular D), and minus if $(D_s, F_s) = (0,1)$ or $(D_s, F_s) = (1,0)$ (i.e., rectangular A or rectangular C).

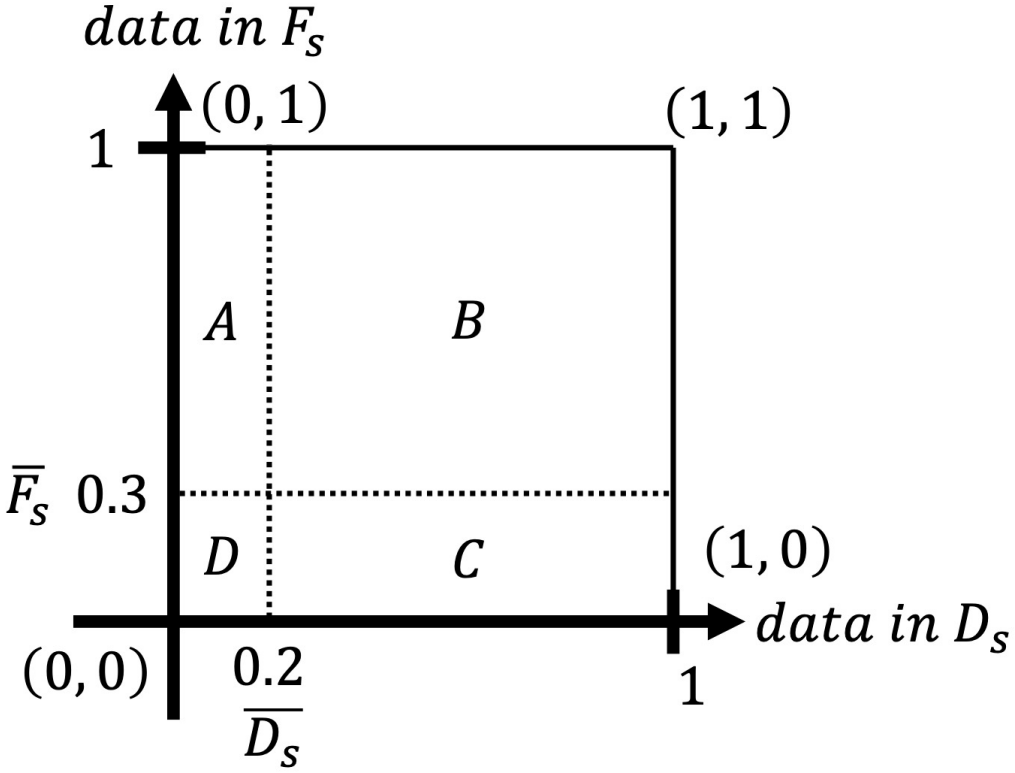

Figure 1. Intuitive image on the bias in correlation coefficient

Notably, the total area of rectangular A and rectangular C is calculated as $|(0-0.2)\cdot(1-0.3)|+|(1-0.2)\cdot(0-0.3)|=0.14+0.24=0.38$, and the total area of rectangular B and rectangular D is calculated as $|(1-0.2)\cdot(1-0.3)|+|(0-0.2)\cdot(0-0.3)|=0.56+0.06=0.62$. Therefore, the numerator of correlation coefficient (i.e., $\sum_{s=1}^{10}\{(D_s-\overline{D_s})\cdot(F_s-\overline{F_s})\}$) is biased as positive value since the total area of rectangular B and rectangular D is larger than that of rectangular A and rectangular C; the sign of the numerator of correlation coefficient is plus in the former and minus in the latter.

Again, the correlation coefficient aims to measure the degree of association between data (i.e., D language and F language in our case). However, $\overline{D_s}$ and $\overline{F_s}$ in Figure 1 will determine only these areas (i.e., the numerator of correlation coefficient) but also the ratio of the areas with plus sign to that with minus sign, which resulted in the numerator of correlation coefficient biased by $\overline{D_s}$ and $\overline{F_s}$ irrelevant to the degree of association between data.

Thus, we cannot apply correlation coefficient to original data in WALS. Since the singular values calculated in previous section (e.g., L1 in Table 11 and L2 in Table 12) are mathematically guaranteed as positive, the obtained vectors are in Euclidean space, which allows us to apply correlation coefficient as meaningful quantities from the viewpoint of statistics.

4. Explanations about multiple linear regression models

In Section 5.3, we have introduced “language as variable” and built multiple linear regression model in the case of Figure1 on the main body that will play significant role in order to select the appropriate dimensions obtained from the coordinates by correspondence analysis:

$$\sqrt{\omega_m}x_{1m} = \beta_{01} + \beta_{21}\sqrt{\omega_m}x_{2m} + \beta_{31}\sqrt{\omega_m}x_{3m} + \beta_{41}\sqrt{\omega_m}x_{4m} + \varepsilon_{1m}$$
$$\sqrt{\omega_m}x_{2m} = \beta_{02} + \beta_{12}\sqrt{\omega_m}x_{1m} + \beta_{32}\sqrt{\omega_m}x_{3m} + \beta_{42}\sqrt{\omega_m}x_{4m} + \varepsilon_{2m}$$
$$\sqrt{\omega_m}x_{3m} = \beta_{03} + \beta_{13}\sqrt{\omega_m}x_{1m} + \beta_{23}\sqrt{\omega_m}x_{2m} + \beta_{43}\sqrt{\omega_m}x_{4m} + \varepsilon_{3m}$$
$$\sqrt{\omega_m}x_{4m} = \beta_{04} + \beta_{14}\sqrt{\omega_m}x_{1m} + \beta_{24}\sqrt{\omega_m}x_{2m} + \beta_{34}\sqrt{\omega_m}x_{3m} + \varepsilon_{4m}$$

$m = 1, 2, \ldots, n-1$, and $i = 1$ (Ainu), $i = 2$ (Chukchi), $i = 3$ (Khalkha), $i = 4$ (Navajo) in Figure1 on the main body.

However, some reader may pose a question what the multiple linear regression model will calculate in the context of linguistics, which is the main focus in this section.

Therefore, the main objective of this section is to illustrate intuitively the multiple linear regression model. Let us start remembering the visual image on "vectors" that have been already introduced and calculated by correspondence analysis in Section 2 on these supplementary materials.

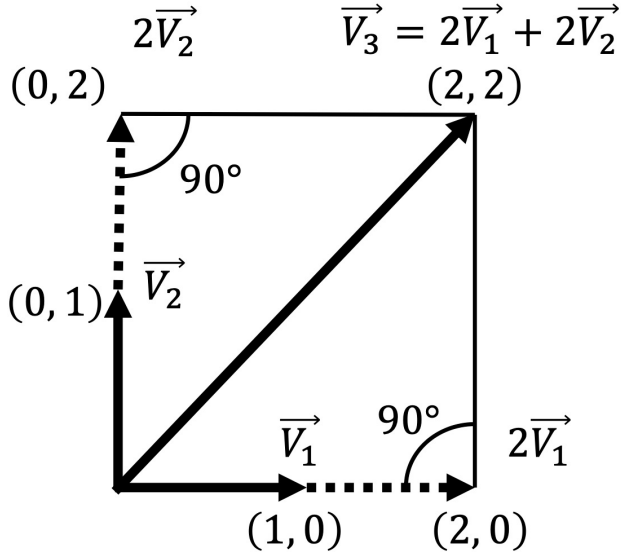

Figure 2. Image on relationships among two-dimensional vectors.

Figure 2 is an image on relationships among two-dimensional vectors. In general, the vector is a series of numbers and represented as $\overrightarrow{V_1}$, $\overrightarrow{V_2}$, and $\overrightarrow{V_3}$ in Figure 2.

For example, $\overrightarrow{V_1} = \binom{1}{0}, \overrightarrow{V_2} = \binom{0}{1}$, and $\overrightarrow{V_3} = \binom{2}{2}$, where the first number of these vector (i.e., 1, 0, and 2) corresponds to the value in horizontal direction in Figure 2, and the second number of the vector (i.e., 0, 1, and 2) corresponds to its value in vertical direction in Figure 2.

In Figure 2, $\overrightarrow{V_3} = 2\overrightarrow{V_1} + 2\overrightarrow{V_2}$, that is $\binom{2}{2} = 2 \cdot \binom{1}{0} + 2 \cdot \binom{0}{1}$. In this case, $\overrightarrow{V_3}$ is totally calculated by the combination of $\overrightarrow{V_1}$ and $\overrightarrow{V_2}$. In this case, we are supposed to define that $\overrightarrow{V_3}$ is totally explained by $\overrightarrow{V_1}$ and $\overrightarrow{V_2}$, which is the fundamental viewpoint in our multiple linear regression model. In other words, $\overrightarrow{V_1}$, $\overrightarrow{V_2}$, and $\overrightarrow{V_3}$ corresponds to the vector of different language in WALS that was calculated by correspondence analysis in in Section 2 on these supplementary materials, respectively. For example, if $\overrightarrow{V_3} = \alpha\overrightarrow{V_1} + \beta\overrightarrow{V_2}$ holds and α and β are real number respectively, then multiple linear regression model will consider that $\overrightarrow{V_3}$ is totally explained by $\overrightarrow{V_1}$ and $\overrightarrow{V_2}$: the language of $\overrightarrow{V_3}$ is totally explained by the language of $\overrightarrow{V_1}$ and the language of $\overrightarrow{V_2}$.

Furthermore, $\overrightarrow{V_2} = \binom{0}{1}$ does not contain any information of $\overrightarrow{V_1} = \binom{1}{0}$ because the first number in $\overrightarrow{V_2}$ (i.e., 0 in $\binom{0}{1}$) cannot vary in $\gamma\overrightarrow{V_2}$ by any real number of γ, and the second number in $\overrightarrow{V_1}$ (i.e., 0 in $\binom{1}{0}$) cannot vary in $\eta\overrightarrow{V_1}$ by any real number of η.

More precisely, we can say that $\overrightarrow{V_1}$ cannot explain $\overrightarrow{V_2}$ or $\overrightarrow{V_2}$ cannot explain $\overrightarrow{V_1}$ because $\overrightarrow{V_1}$ and $\overrightarrow{V_2}$ are in "right angle".

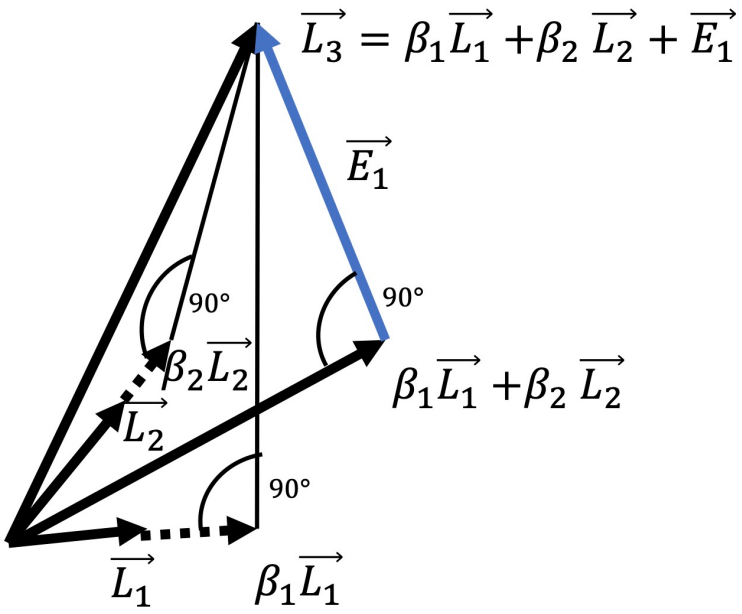

Figure 3. Image on relationships among three-dimensional vectors.

Let us consider more complicated example in Figure 3. $\overrightarrow{L_1}$, $\overrightarrow{L_2}$, and $\overrightarrow{L_3}$ are assumed to be the vectors of the corresponding language in WALS that was calculated by correspondence analysis in in Section 2. In this case, $\overrightarrow{L_3} = \alpha\overrightarrow{L_1} + \beta\overrightarrow{L_2}$ does not hold in any real numbers of α and β, and $\overrightarrow{L_3}$ is not totally explained by $\overrightarrow{L_1}$ and $\overrightarrow{L_2}$. Our alternative is to decompose $\overrightarrow{L_3}$ into two parts: one that is totally explained by $\overrightarrow{L_1}$ and $\overrightarrow{L_2}$, and the other that cannot explained by either $\overrightarrow{L_1}$ or $\overrightarrow{L_2}$ at all. Therefore, we will calculate $\beta_1\overrightarrow{L_1} + \beta_2\overrightarrow{L_2}$ as $\beta_1\overrightarrow{L_1} + \beta_2\overrightarrow{L_2}$ and $\overrightarrow{L_3} - \beta_1\overrightarrow{L_1} + \beta_2\overrightarrow{L_2}$ are in right angle as explained in Figure 2 in our supplementary materials. Furthermore, the additional restriction requires us minimizing the length of $\overrightarrow{L_3} - \beta_1\overrightarrow{L_1} + \beta_2\overrightarrow{L_2}$, the departure from $\overrightarrow{L_3}$ to $\beta_1\overrightarrow{L_1} + \beta_2\overrightarrow{L_2}$ corresponding to the information in $\overrightarrow{L_3}$ not explained by $\beta_1\overrightarrow{L_1} + \beta_2\overrightarrow{L_2}$.

Thus, we can calculate β_1 *and* β_2 in $\beta_1\overrightarrow{L_1} + \beta_2\overrightarrow{L_2}$ and decompose $\overrightarrow{L_3}$ into two parts. Again, one is totally explained by $\overrightarrow{L_1}$ and $\overrightarrow{L_2}$, and the other cannot by explained by either $\overrightarrow{L_1}$ or $\overrightarrow{L_2}$ at all.

Let us return to our multiple linear regression model:

$$\sqrt{\omega_m}x_{1m} = \beta_{01} + \beta_{21}\sqrt{\omega_m}x_{2m} + \beta_{31}\sqrt{\omega_m}x_{3m} + \beta_{41}\sqrt{\omega_m}x_{4m} + \varepsilon_{1m}$$
$$\sqrt{\omega_m}x_{2m} = \beta_{02} + \beta_{12}\sqrt{\omega_m}x_{1m} + \beta_{32}\sqrt{\omega_m}x_{3m} + \beta_{42}\sqrt{\omega_m}x_{4m} + \varepsilon_{2m}$$
$$\sqrt{\omega_m}x_{3m} = \beta_{03} + \beta_{13}\sqrt{\omega_m}x_{1m} + \beta_{23}\sqrt{\omega_m}x_{2m} + \beta_{43}\sqrt{\omega_m}x_{4m} + \varepsilon_{3m}$$
$$\sqrt{\omega_m}x_{4m} = \beta_{04} + \beta_{14}\sqrt{\omega_m}x_{1m} + \beta_{24}\sqrt{\omega_m}x_{2m} + \beta_{34}\sqrt{\omega_m}x_{3m} + \varepsilon_{4m}$$

In the first equation, $\sqrt{\omega_m}x_{1m} = \beta_{01} + \beta_{21}\sqrt{\omega_m}x_{2m} + \beta_{31}\sqrt{\omega_m}x_{3m} + \beta_{41}\sqrt{\omega_m}x_{4m} + \varepsilon_{1m}$, $\sqrt{\omega_m}x_{1m}$ corresponds to the m-dimensional vector of the first language (e.g., Ainu in Figure 1 on the main body), $\sqrt{\omega_m}x_{2m}$ to the m-dimensional vector of the second language (e.g., Chukchi in Figure 1 on the main body), $\sqrt{\omega_m}x_{3m}$ to the m-dimensional vector of the third language (e.g., Khalkha in Figure 1 on the main body), $\sqrt{\omega_m}x_{4m}$ to the m-dimensional vector of the fourth language (e.g., Navajo in Figure 1 on the main body), and these m-dimensional vectors are calculated by correspondence analysis. In this case, ε_{1m} corresponds to the m-dimensional vector that cannot explained by $\sqrt{\omega_m}x_{1m}$, $\sqrt{\omega_m}x_{2m}$, $\sqrt{\omega_m}x_{3m}$, and $\sqrt{\omega_m}x_{4m}$, and ε_{1m} and $\beta_{01} + \beta_{21}\sqrt{\omega_m}x_{2m} + \beta_{31}\sqrt{\omega_m}x_{3m} + \beta_{41}\sqrt{\omega_m}x_{4m}$ are in right angle as shown in the example of Figure 3 here. The same analysis can apply to other equation in multiple linear regression model.

Thus, the m-dimensional vectors of ε_{1m}, ε_{2m}, ε_{3m}, ε_{4m} are the information of the concerned language cannot be explained by the other languages as explained in Section 5.3 on the main body.

Notably, in Section 5.3, we proposed to measure correlation coefficient among the m-dimensional vectors of ε_{1m}, ε_{2m}, ε_{3m}, ε_{4m} as the similarity between two languages not explained by the other languages, which can be considered as the similarity between languages not explained by linguistic typology on the other languages but explained by other factors than linguistic typology between the two languages (e.g., geographical factor) as explained in Section 5.3.

Since linguists often explain one language by other languages and state "one language is similar to the group of other languages" (e.g., linguists say "English is the Indo-European language family or similar to languages in Indo-European language family"), this viewpoint in our multiple linear regression model can be considered as some way of thinking that areal linguists will capture the areal information or relationships in languages or historical linguistics will reconstruct the genealogical information or relationships in languages, I have adopted multiple linear regression model in this article. Thus, further investigations are promising whether the viewpoint can adequately capture areal and genealogical information in linguistic typology, and whether there is an alternative to improve the viewpoint in multiple linear regression appropriate in the context of linguistic typology.

5. Results

The detail results omitted in the main body of this paper are in Figures 4-15, and Tables 15-20.

6. WALS data

The World Atlas of Language Structures Online (abbreviate as WALS hereafter) is a linguistic database constructed by a team of 55 linguists (most of them leading authorities in the relevant subfield), organized around various linguistic parameters (referred to in WALS as features). WALS contains 144 chapters, each consisting of a text and a main map. Each of the 144 chapters shows the distribution of a particular linguistic feature, reflected in the chapter's title. In several cases, a single chapter includes more than one map. Most WALS features correspond straightforwardly to chapters, but some chapters describe multiple features.

Oxford University Press published the first version of WALS as a book with an accompanying CD-ROM in 2005. The first online version was published in April 2008. Both are superseded by the

current online version, released in April 2011. WALS is a joint effort of the Max Planck Institute for Evolutionary Anthropology and the Max Planck Digital Library (For a detailed explanation, see http://wals.info/). The list of languages and features is available at wals.info/languoid (http://wals.info/languoid) and the list of features at wals.info/feature (see http://wals.info/feature). WALS provides data on 2,676 languages and 192 linguistic features constructed by domain: phonology, word order, lexicon, word order, nominal categories, etc. Utilizing statistical analysis explained in Sections 2 and 5 requires transforming the original categorical data into binary data. The total number of values associated with the 192 features gives some 1200 categories (feature values). We reduced the number of these categories to 489 to eliminate those categories for which there were less than 10 applicable languages. Our sample of 201 languages was chosen based on the recommended WALS 200 language sample (http://wals.info/languoid/samples/200), which is balanced by genetic affiliation and region. Note that we have removed the Muong language containing too much NA in WALS from our data.

We call the database consisting of 489 values and 201 languages except Muong as "Data 1" used in Whitman and Ono (2017), those removing all word-order parameters (81A-97A, 90A-90G and 143A-144Y) from Data 1 as "Data 2", those removing parameters (81A-86A, 88A-90A, 90A-90G, 94A, 95A, 143A-144Y), which Whitman and Ono (2017) identified as the main component, as "Data 3", with the parameters on adjective and negation (87A[Order of Adjective and Noun], 143A[Order of Negative Morpheme and Verb], 143E[Preverbal Negative Morphemes], 143F[Postverbal Negative Morphemes], 143G[Minor morphological means], 144B[Position of negative words relative to beginning and end of clause and with respect to adjacency to verb], 144V[Verb-initial with Preverbal Negative], 144X[Verb-initial with Clause-Final Negative]) that Dryer (1992) indicated, retained. Note that we excluded parameters related to the Order of Objects and Nouns (e.g., 144S) because such parameters are redundant and strongly correlated. We avoid biasing the result of clustering, removing those parameters. Data 2 contains 391 values and 201 languages, and Data 3 has 429 values and 201 languages. Data 1, Data 2, and Data 3 are the same data used in Whitman and Ono (2015) and Ono and Whitman (2016).

Furthermore, Data 4 is comprised of 203 languages in WALS data updated (Dryer and Haspelmath 2013b) that have revised Bunuba to Malakmalak and added Burarra and Sedang to Data 1-3, removing those categories for which there were less than 10 applicable languages and all word-order except Adjective and Negation as Data 3. Thus, Data 4 consists of 439 values and 203 languages.

Table 17, Table 18, Table 19, and Table 20 correspond to features and values in Data 1, Data 2, Data 3, and Data 4, respectively. The details explanations can refer to Dryer and Haspelmath (2013a) about Data 1, Data 2 and Data 3, and to Dryer and Haspelmath (2013b) about Data 4.

References

Dryer, Matthew S. and Martin Haspelmath (eds.) (2013a). The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2014-07-28.)

Dryer, Matthew S. and Martin Haspelmath (eds.) (2013b). The World Atlas of Language Structures

Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2018-03-09.)

Greenacre, Michael (2017) *Correspondence analysis in practice*. CRC Press.

Maddieson, Ian (2013a) Consonant inventories. In Dryer, Matthew S. and Martin Haspelmath (eds.), The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/14, Accessed on 2014-07-28.)

Maddieson, Ian (2013b) Consonant inventories. In Dryer, Matthew S. and Martin Haspelmath (eds.), The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/14, Accessed on 2018-03-09.)

Kimiyama Yoshirō (2011) *Koresupondensu bunseki no riyou houhou: ippanka taiou bunseki moderu [Applications on correspondence analysis: general correspondence analysis]* third edition, Tokyo: Data Analysis Institute, Inc.

Table 15. The results of Rand Index in each clustering and data condition.

Data	Method	Grouping	Distance	Rand_Index
Data3	MCA-full-cluster	ward	correlation	0.8908955220
Data3	Cluster Analysis	ward	correlation	0.8896517410
Data1	Cluster Analysis	ward	correlation	0.8895522390
Data2	MCA-full-cluster	complete	correlation	0.8895522390
Data2	MCA-full-cluster	ward	correlation	0.8889054730
Data1	MCA-full-cluster	ward	correlation	0.8885074630
Data2	Cluster Analysis	complete	correlation	0.8880099500
Data3	Cluster Analysis	complete	correlation	0.8871144280
Data1	MCA-full-cluster	complete	correlation	0.8864179100
Data2	Cluster Analysis	ward	correlation	0.8853233830
Data1	Cluster Analysis	complete	manhattan	0.8843781090
Data1	MCA-tandem	ward	manhattan	0.8838308460
Data3	Cluster Analysis	complete	manhattan	0.8837810950
Data2	Cluster Analysis	complete	manhattan	0.8824875620
Data3	MCA-full-cluster	complete	correlation	0.8802487560
Data1	Cluster Analysis	ward	manhattan	0.8799502490
Data3	Cluster Analysis	ward	euclidean	0.8794029850
Data2	Cluster Analysis	ward	euclidean	0.8793034830
Data1	MCA-full-cluster	ward	manhattan	0.8788059700
Data1	MCA-tandem	ward	euclidean	0.8786567160
Data1	Cluster Analysis	complete	euclidean	0.8779104480
Data3	MCA-full-cluster	ward	euclidean	0.8766169150
Data1	MCA-tandem	complete	manhattan	0.8761691540
Data3	Cluster Analysis	ward	manhattan	0.8741791040
Data2	MCA-full-cluster	ward	euclidean	0.8731840800
Data2	Cluster Analysis	ward	manhattan	0.8730348260
Data3	MCA-tandem	complete	euclidean	0.8716915420
Data1	MCA-full-cluster	ward	euclidean	0.8705970150
Data1	Cluster Analysis	complete	correlation	0.8701492540
Data1	Cluster Analysis	ward	euclidean	0.8693034830
Data3	MCA-tandem	complete	manhattan	0.8693034830
Data2	MCA-tandem	ward	manhattan	0.8692039800
Data3	MCA-tandem	ward	manhattan	0.8692039800
Data2	MCA-tandem	ward	euclidean	0.8680099500
Data1	MCA-tandem	complete	euclidean	0.8648756220
Data2	Cluster Analysis	complete	euclidean	0.8618905470
Data3	MCA-tandem	ward	euclidean	0.8614427860
Data2	MCA-tandem	complete	manhattan	0.8595024880
Data2	MCA-tandem	complete	euclidean	0.8572139300
Data3	Cluster Analysis	complete	euclidean	0.8534825870
Data3	MCA-full-cluster	ward	manhattan	0.8350746270
Data2	MCA-full-cluster	ward	manhattan	0.8281094530
Data1	MCA-full-cluster	complete	euclidean	NA
Data2	MCA-full-cluster	complete	euclidean	NA
Data3	MCA-full-cluster	complete	euclidean	NA
Data1	MCA-full-cluster	complete	manhattan	NA
Data2	MCA-full-cluster	complete	manhattan	NA
Data3	MCA-full-cluster	complete	manhattan	NA
Data1	MCA-tandem	ward	correlation	NA
Data2	MCA-tandem	ward	correlation	NA
Data3	MCA-tandem	ward	correlation	NA

Table 16. The results by Partitioning Around Method (PAM)

Name	genus	family	macroarea	C1_id	C2_id	C3_id	C4_id	C5_id	C6_id
Ainu	Ainu	Ainu	Eurasia	1	7	1	1	3	3
Abkhaz	Northwest Caucasian	Northwest Caucasian	Eurasia	1	1	1	5	1	2
Ket	Yeniseian	Yeniseian	Eurasia	1	5	1	1	5	5
Cree (Plains)	Algonquian	Algic	North America	1	1	1	1	1	4
Yurok	Yurok	Algic	North America	1	1	1	2	1	3
Wichita	Caddoan	Caddoan	North America	1	1	1	1	7	2
Haida	Haida	Haida	North America	1	3	1	2	3	7
Maricopa	Yuman	Hokan	North America	1	3	1	2	5	5
Oneida	Northern Iroquoian	Iroquoian	North America	1	1	1	1	4	3
Karok	Karok	Karok	North America	1	1	1	4	1	7
Acoma	Keresan	Keresan	North America	1	1	1	1	2	1
Kiowa	Kiowa-Tanoan	Kiowa-Tanoan	North America	1	2	1	1	2	7
Kutenai	Kutenai	Kutenai	North America	1	1	1	1	7	1
Jakaltek	Mayan	Mayan	North America	1	1	1	1	6	2
Koasati	Muskogean	Muskogean	North America	1	1	1	1	4	7
Navajo	Athapaskan	Na-Dene	North America	1	1	1	1	5	6
Slave	Athapaskan	Na-Dene	North America	1	1	1	4	5	6
Tlingit	Tlingit	Na-Dene	North America	1	1	1	1	3	1
Coos (Hanis)	Coosan	Oregon Coast	North America	1	1	1	2	2	3
Otomí (Mezquital)	Otomian	Oto-Manguean	North America	1	1	1	3	3	2
Nez Perce	Sahaptian	Penutian	North America	1	1	1	2	6	6
Tsimshian (Coast)	Tsimshianic	Penutian	North America	1	1	1	3	4	5
Squamish	Central Salish	Salishan	North America	1	1	1	1	7	4
Lakhota	Core Siouan	Siouan	North America	1	1	1	1	2	2
Makah	Southern Wakashan	Wakashan	North America	1	3	1	2	6	1
Yuchi	Yuchi	Yuchi	North America	1	1	1	1	3	6
Mapudungun	Araucanian	Araucanian	South America	1	7	1	4	6	5
Apurinã	Purus	Arawakan	South America	1	1	1	1	4	3
Abipón	South Guaicuruan	Guaicuruan	South America	1	1	1	3	1	1
Wichí	Matacoan	Matacoan	South America	1	2	1	1	1	1
Yagua	Peba-Yaguan	Peba-Yaguan	South America	1	1	1	1	7	3
Trumai	Trumai	Trumai	South America	1	3	5	2	4	5
Guaraní	Tupi-Guaraní	Tupian	South America	1	1	1	4	4	7
Gooniyandi	Bunuban	Bunuban	Australia	2	2	2	2	3	6
Nunggubuyu	Nunggubuyu	Gunwinyguan	Australia	2	2	2	2	3	7
Maung	Iwaidjan	Iwaidjan	Australia	2	2	2	4	1	4
Mangarrayi	Mangarrayi	Mangarrayi-Maran	Australia	2	2	2	2	7	4
Burarra	Burarran	Mangrida	Australia	2	1	2	2	7	1
Wambaya	Wambayan	Mirndi	Australia	2	2	2	2	3	6
Malakmalak	Northern Daly	Northern Daly	Australia	2	2	2	2	6	2
Martuthunira	Western Pama-Nyungan	Pama-Nyungan	Australia	2	5	2	5	6	2
Ngiyambaa	Southeastern Pama-Nyungan	Pama-Nyungan	Australia	2	5	2	2	7	1
Pitjantjatjara	Western Pama-Nyungan	Pama-Nyungan	Australia	2	2	2	5	1	2
Yidiny	Northern Pama-Nyungan	Pama-Nyungan	Australia	2	5	2	2	2	6
Kayardild	Tangkic	Tangkic	Australia	2	3	2	5	7	6
Tiwi	Tiwian	Tiwian	Australia	2	2	2	1	7	7
Maranungku	Wagaydy	Western Daly	Australia	2	2	2	1	5	3
Ungarinjin	Worrorran	Worrorran	Australia	2	2	2	2	5	6
Wardaman	Yangmanic	Yangmanic	Australia	2	2	2	2	1	3
Chukchi	Northern Chukotko-Kamchatkan	Chukotko-Kamchatkan	Eurasia	2	3	6	5	3	2
Passamaquoddy-Maliseet	Algonquian	Algic	North America	2	6	4	1	5	7
Rama	Rama	Chibchan	North America	2	2	4	5	5	2
Yup'ik (Central)	Eskimo	Eskimo-Aleut	North America	2	2	6	2	2	6
Zoque (Copainalá)	Mixe-Zoque	Mixe-Zoque	North America	2	2	2	1	5	4
Miwok (Southern Sierra)	Miwok	Penutian	North America	2	2	2	2	6	1
Cahuilla	California Uto-Aztecan	Uto-Aztecan	North America	2	2	2	2	2	5
Comanche	Numic	Uto-Aztecan	North America	2	3	2	1	4	3
Yaqui	Cahita	Uto-Aztecan	North America	2	2	2	5	4	4
Imonda	Border	Border	Papunesia	2	2	2	5	5	4
Yimas	Lower Sepik	Lower Sepik-Ramu	Papunesia	2	6	2	4	6	4
Alamblak	Sepik Hill	Sepik	Papunesia	2	6	4	1	3	4
Arapesh (Mountain)	Kombio-Arapesh	Torricelli	Papunesia	2	2	2	4	5	1
Paumarí	Arauan	Arauan	South America	2	3	2	1	6	4
Bawm	Kuki-Chin	Sino-Tibetan	Eurasia	3	3	5	3	1	5
Bribri	Talamanca	Chibchan	North America	3	3	5	5	4	4
Tunica	Tunica	Tunica	North America	3	1	4	1	6	2
Nahuatl (Tetelcingo)	Aztecan	Uto-Aztecan	North America	3	1	3	2	5	6
Daga	Dagan	Dagan	Papunesia	3	1	3	1	7	1
Marind	Marind Proper	Marind	Papunesia	3	6	3	4	6	7

(Continue)

Name	genus	family	macroarea	C1_id	C2_id	C3_id	C4_id	C5_id	C6_id
Sentani	Sentani	Sentani	Papunesia	3	6	3	3	1	1
Lavukaleve	Lavukaleve	Solomons East Papuan	Papunesia	3	2	3	4	4	3
Asmat	Asmat-Kamoro	Trans-New Guinea	Papunesia	3	7	3	7	2	2
Dani (Lower Grand Valley)	Dani	Trans-New Guinea	Papunesia	3	3	3	3	2	4
Ekari	Wissel Lakes-Kemandoga	Trans-New Guinea	Papunesia	3	3	3	4	4	7
Hamtai	Angan	Trans-New Guinea	Papunesia	3	6	3	3	4	3
Kewa	Engan	Trans-New Guinea	Papunesia	3	5	3	4	5	4
Kobon	Madang	Trans-New Guinea	Papunesia	3	2	3	2	6	2
Suena	Binanderean	Trans-New Guinea	Papunesia	3	3	3	5	7	4
Una	Mek	Trans-New Guinea	Papunesia	3	1	5	3	2	4
Usan	Madang	Trans-New Guinea	Papunesia	3	1	3	1	7	3
Carib	Cariban	Cariban	South America	3	6	3	3	5	4
Hixkaryana	Cariban	Cariban	South America	3	6	3	4	3	5
Cayuvava	Cayuvava	Cayuvava	South America	3	2	3	1	2	4
Ika	Arhuacic	Chibchan	South America	3	5	3	3	4	5
Canela-Krahô	Ge-Kaingang	Macro-Ge	South America	3	7	3	7	1	2
Pirahã	Mura	Mura	South America	3	3	3	7	1	3
Urubú-Kaapor	Tupi-Guaraní	Tupian	South America	3	6	3	3	3	7
Warao	Warao	Warao	South America	3	3	3	5	2	2
Sanuma	Yanomam	Yanomam	South America	3	3	3	2	6	4
Arabic (Egyptian)	Semitic	Afro-Asiatic	Africa	4	4	4	6	1	3
Beja	Beja	Afro-Asiatic	Africa	4	1	4	3	5	1
Berber (Middle Atlas)	Berber	Afro-Asiatic	Africa	4	4	4	6	1	5
Hausa	West Chadic	Afro-Asiatic	Africa	4	6	4	3	3	5
Iraqw	Southern Cushitic	Afro-Asiatic	Africa	4	4	4	3	5	5
Kera	East Chadic	Afro-Asiatic	Africa	4	6	5	7	3	7
Oromo (Harar)	Lowland East Cushitic	Afro-Asiatic	Africa	4	4	4	4	3	7
Bagirmi	Bongo-Bagirmi	Central Sudanic	Africa	4	7	7	7	5	7
Ngiti	Lendu	Central Sudanic	Africa	4	1	7	2	1	7
Lango	Nilotic	Eastern Sudanic	Africa	4	7	7	4	7	2
Murle	Surmic	Eastern Sudanic	Africa	4	7	7	7	7	3
Fur	Fur	Fur	Africa	4	3	4	3	3	7
Krongo	Kadugli	Kadu	Africa	4	3	4	4	4	2
Kunama	Kunama	Kunama	Africa	4	5	4	3	7	5
Maba	Maban	Maban	Africa	4	3	4	3	4	3
Bambara	Western Mande	Mande	Africa	4	3	5	3	3	1
Diola-Fogny	Northern Atlantic	Niger-Congo	Africa	4	6	4	4	2	3
Ewe	Kwa	Niger-Congo	Africa	4	6	4	4	7	1
Grebo	Kru	Niger-Congo	Africa	4	6	7	4	4	6
Igbo	Igboid	Niger-Congo	Africa	4	7	5	3	2	3
Kongo	Bantoid	Niger-Congo	Africa	4	6	3	4	6	6
Luvale	Bantoid	Niger-Congo	Africa	4	6	3	4	7	6
Nkore-Kiga	Bantoid	Niger-Congo	Africa	4	6	3	4	1	6
Sango	Ubangi	Niger-Congo	Africa	4	7	7	7	4	4
Supyire	Gur	Niger-Congo	Africa	4	6	5	4	1	4
Swahili	Bantoid	Niger-Congo	Africa	4	6	6	4	1	6
Zulu	Bantoid	Niger-Congo	Africa	4	6	3	4	5	6
Kanuri	Western Saharan	Saharan	Africa	4	5	4	3	7	2
Koyraboro Senni	Songhay	Songhay	Africa	4	7	4	3	4	1
Lepcha	Lepcha	Sino-Tibetan	Eurasia	4	3	5	3	2	7
Paamese	Oceanic	Austronesian	Papunesia	4	6	4	3	6	6
Amele	Madang	Trans-New Guinea	Papunesia	4	5	4	3	3	2
Nubian (Dongolese)	Nubian	Eastern Sudanic	Africa	5	5	4	3	1	5
Evenki	Tungusic	Altaic	Eurasia	5	5	6	5	2	7
Khalkha	Mongolic	Altaic	Eurasia	5	5	5	5	5	7
Turkish	Turkic	Altaic	Eurasia	5	5	6	5	3	7
Mundari	Munda	Austro-Asiatic	Eurasia	5	5	5	1	2	7
Burushaski	Burushaski	Burushaski	Eurasia	5	1	6	5	3	4
Brahui	Northern Dravidian	Dravidian	Eurasia	5	5	6	5	6	2
Kannada	Southern Dravidian	Dravidian	Eurasia	5	5	6	5	2	4
Greenlandic (West)	Eskimo	Eskimo-Aleut	Eurasia	5	3	6	5	4	2
Japanese	Japanese	Japanese	Eurasia	5	5	5	5	2	6
Korean	Korean	Korean	Eurasia	5	5	5	5	6	6
Hunzib	Avar-Andic-Tsezic	Nakh-Daghestanian	Eurasia	5	5	5	5	1	7
Ingush	Nakh	Nakh-Daghestanian	Eurasia	5	5	5	5	2	2
Lak	Lak-Dargwa	Nakh-Daghestanian	Eurasia	5	1	5	5	5	3
Lezgian	Lezgic	Nakh-Daghestanian	Eurasia	5	5	5	5	7	2
Nivkh	Nivkh	Nivkh	Eurasia	5	5	5	5	1	3
Burmese	Burmese-Lolo	Sino-Tibetan	Eurasia	5	5	5	7	1	7

(Continue)

Name	genus	family	macroarea	C1_id	C2_id	C3_id	C4_id	C5_id	C6_id
Garo	Bodo-Garo	Sino-Tibetan	Eurasia	5	5	5	5	3	3
Ladakhi	Bodic	Sino-Tibetan	Eurasia	5	3	5	5	2	7
Meithei	Kuki-Chin	Sino-Tibetan	Eurasia	5	5	5	5	4	1
Yukaghir (Kolyma)	Yukaghir	Yukaghir	Eurasia	5	5	6	5	1	3
Pomo (Southeastern)	Pomoan	Hokan	North America	5	3	5	5	7	5
Qawasqar	Alacalufan	Alacalufan	South America	5	3	5	3	3	1
Aymara (Central)	Aymaran	Aymaran	South America	5	2	6	5	4	5
Awa Pit	Barbacoan	Barbacoan	South America	5	3	6	4	4	1
Epena Pedee	Choco	Choco	South America	5	3	5	5	2	4
Selknam	Chon Proper	Chon	South America	5	1	5	1	6	5
Huitoto (Minica)	Huitoto	Huitotoan	South America	5	5	5	3	5	6
Shipibo-Konibo	Panoan	Panoan	South America	5	3	5	5	5	1
Quechua (Imbabura)	Quechuan	Quechuan	South America	5	5	6	5	2	3
Araona	Tacanan	Tacanan	South America	5	3	5	3	5	5
Barasano	Tucanoan	Tucanoan	South America	5	2	6	2	5	1
Khoekhoe	Khoe-Kwadi	Khoe-Kwadi	Africa	6	2	2	1	7	7
Hebrew (Modern)	Semitic	Afro-Asiatic	Eurasia	6	4	6	6	7	5
Basque	Basque	Basque	Eurasia	6	4	6	5	5	2
Armenian (Eastern)	Armenian	Indo-European	Eurasia	6	4	6	6	3	6
English	Germanic	Indo-European	Eurasia	6	4	6	6	2	1
French	Romance	Indo-European	Eurasia	6	4	6	6	6	3
German	Germanic	Indo-European	Eurasia	6	4	6	6	3	3
Greek (Modern)	Greek	Indo-European	Eurasia	6	4	6	6	1	4
Hindi	Indic	Indo-European	Eurasia	6	4	6	6	2	5
Irish	Celtic	Indo-European	Eurasia	6	4	6	6	1	2
Latvian	Baltic	Indo-European	Eurasia	6	4	6	6	4	3
Persian	Iranian	Indo-European	Eurasia	6	4	6	6	5	4
Russian	Slavic	Indo-European	Eurasia	6	4	6	6	6	5
Spanish	Romance	Indo-European	Eurasia	6	4	6	6	2	2
Georgian	Kartvelian	Kartvelian	Eurasia	6	4	6	6	5	5
Finnish	Finnic	Uralic	Eurasia	6	4	6	6	4	7
Hungarian	Ugric	Uralic	Eurasia	6	4	6	6	4	3
Nenets	Samoyedic	Uralic	Eurasia	6	5	6	6	2	7
Malagasy	Barito	Austronesian	Africa	7	7	7	7	3	6
Ju|'hoan	Ju-Kung	Kxa	Africa	7	6	7	7	6	4
Koromfe	Gur	Niger-Congo	Africa	7	6	7	3	3	2
Yoruba	Defoid	Niger-Congo	Africa	7	7	7	7	1	4
Khasi	Khasian	Austro-Asiatic	Eurasia	7	7	7	7	2	7
Khmer	Khmer	Austro-Asiatic	Eurasia	7	7	7	7	7	1
Khmu'	Palaung-Khmuic	Austro-Asiatic	Eurasia	7	7	7	7	2	3
Sedang	Bahnaric	Austro-Asiatic	Eurasia	7	7	7	7	3	5
Semelai	Aslian	Austro-Asiatic	Eurasia	7	7	7	7	1	5
Vietnamese	Viet-Muong	Austro-Asiatic	Eurasia	7	7	7	7	3	4
Hmong Njua	Hmong-Mien	Hmong-Mien	Eurasia	7	7	7	7	3	2
Kayah Li (Eastern)	Karen	Sino-Tibetan	Eurasia	7	7	7	7	7	6
Mandarin	Chinese	Sino-Tibetan	Eurasia	7	7	7	7	6	6
Thai	Kam-Tai	Tai-Kadai	Eurasia	7	7	7	7	1	1
Chinantec (Lealao)	Chinantecan	Oto-Manguean	North America	7	6	7	2	3	6
Mixtec (Chalcatongo)	Mixtecan	Oto-Manguean	North America	7	7	7	7	5	6
Batak (Karo)	Northwest Sumatra-Barrier Islands	Austronesian	Papunesia	7	7	7	7	6	7
Chamorro	Chamorro	Austronesian	Papunesia	7	7	7	1	7	4
Drehu	Oceanic	Austronesian	Papunesia	7	7	7	7	7	6
Fijian	Oceanic	Austronesian	Papunesia	7	7	7	7	4	5
Indonesian	Malayo-Sumbawan	Austronesian	Papunesia	7	7	7	7	4	4
Kilivila	Oceanic	Austronesian	Papunesia	7	6	7	7	4	7
Kiribati	Oceanic	Austronesian	Papunesia	7	6	7	7	4	5
Maori	Oceanic	Austronesian	Papunesia	7	7	7	7	5	2
Paiwan	Paiwan	Austronesian	Papunesia	7	2	7	2	7	4
Rapanui	Oceanic	Austronesian	Papunesia	7	7	7	7	6	3
Taba	South Halmahera - West New Guinea	Austronesian	Papunesia	7	7	7	7	1	7
Tagalog	Greater Central Philippine	Austronesian	Papunesia	7	7	7	7	2	7
Tukang Besi	Celebic	Austronesian	Papunesia	7	7	7	7	4	4
Maybrat	North-Central Bird's Head	West Papuan	Papunesia	7	6	7	1	3	4
Wari'	Chapacura-Wanham	Chapacura-Wanham	South America	7	2	7	7	3	5
Ndyuka	Creoles and Pidgins	other	South America	7	4	7	6	4	6

C1: From 1st to 12st dimension and from 146st to 202nd dimensions; C2: all dimensions (from 1st to 202nd dimension); C3: From 1st to 12st dimension; C4: From 1st to 145st dimension; C5: From 146st to 202nd dimension; C6: From 13st to 202nd dimension.

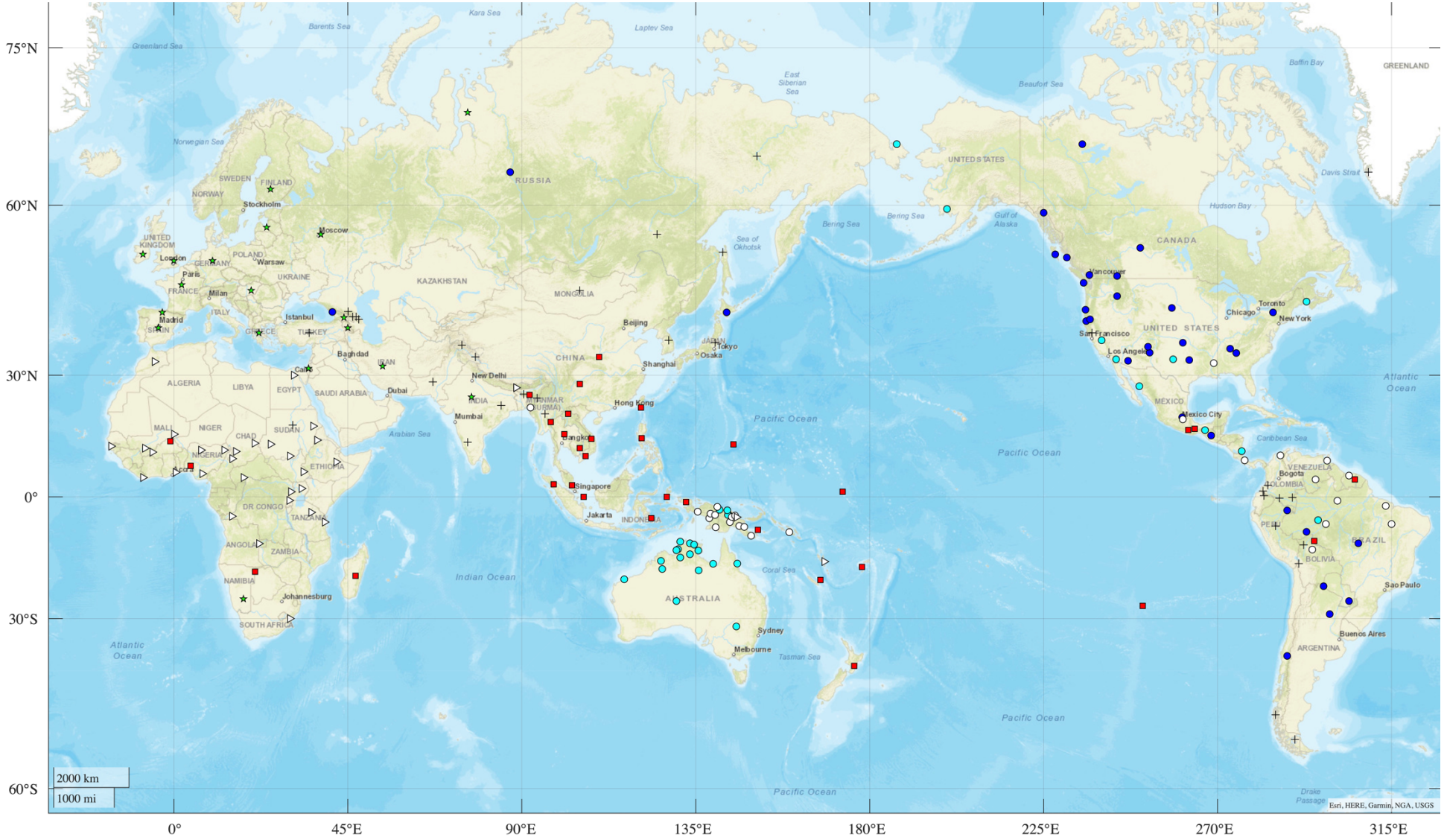

Figure 4: The classification results in C1 (i.e., from 1st to 12th dimension and from the 146th to 202nd dimensions).

Blue circle: cluster 1; cyan circle: cluster 2; white circle: cluster 3; triangle: cluster 4; plus: cluster 5; star: cluster 6; rectangle: cluster 7 in C1 in Table 16.

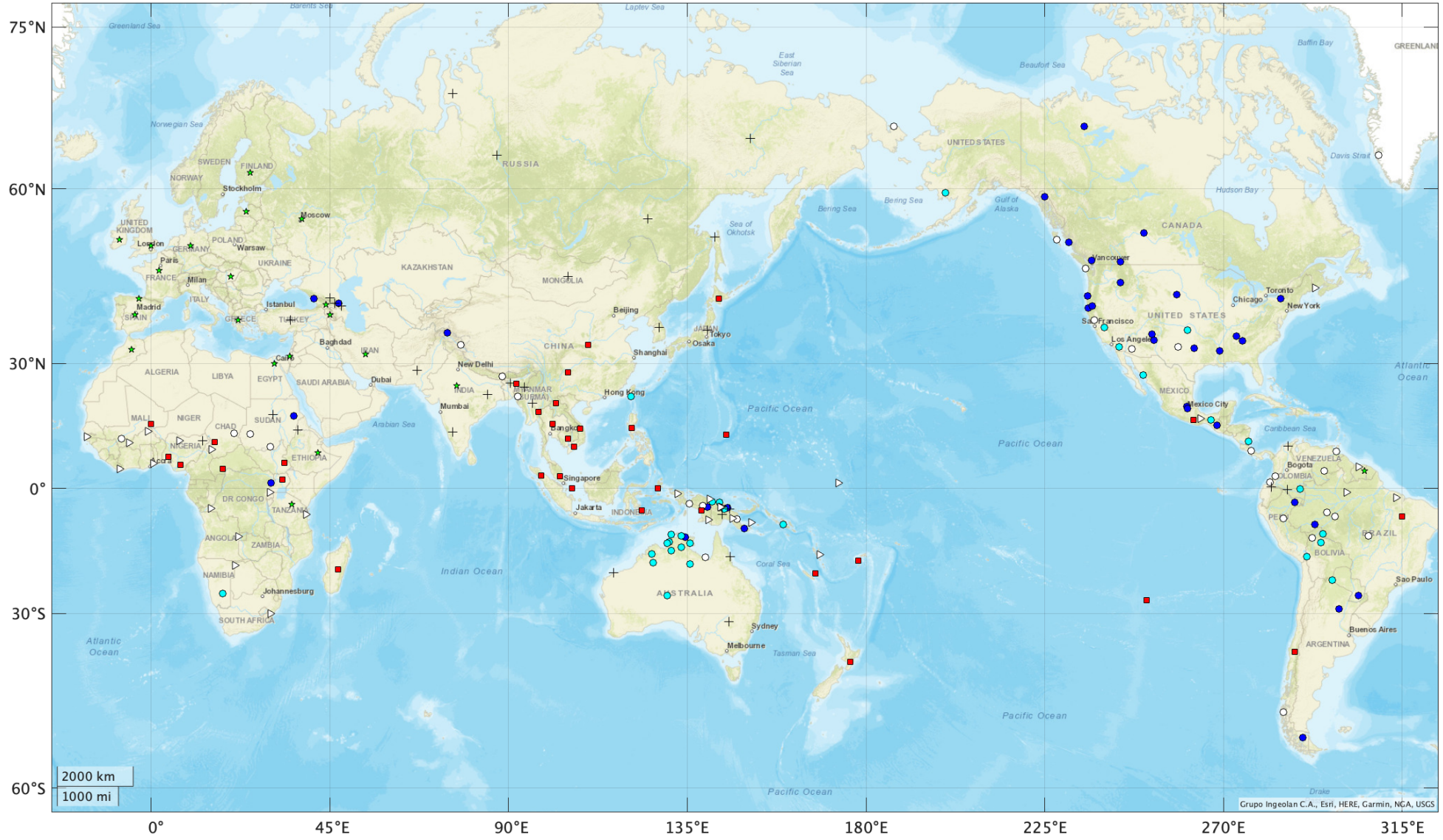

Figure 5: The classification results in C2 (i.e., from 1st to 202nd dimensions).

Blue circle: cluster 1; cyan circle: cluster 2; white circle: cluster 3; triangle: cluster 6; plus: cluster 5; star: cluster 4; rectangle: cluster 7 in C 2 in Table 16.

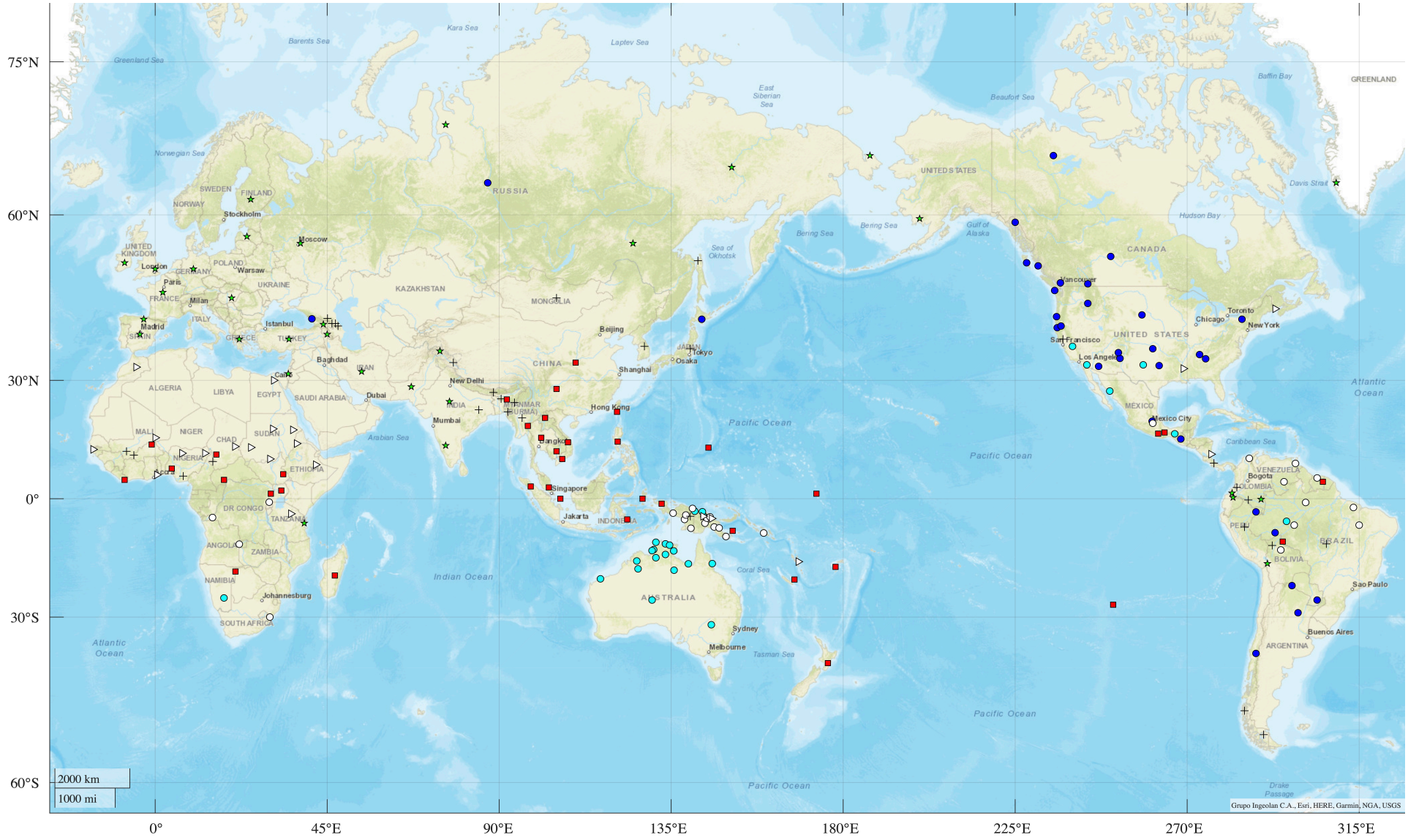

Figure 6: The classification results in C3 (i.e., from the 1st to the 12st dimension).

Blue circle: cluster 1; cyan circle: cluster 2; white circle: cluster 3; triangle: cluster 4; plus: cluster 5; star: cluster 6; rectangle: cluster 7 in C3 in Table 16.

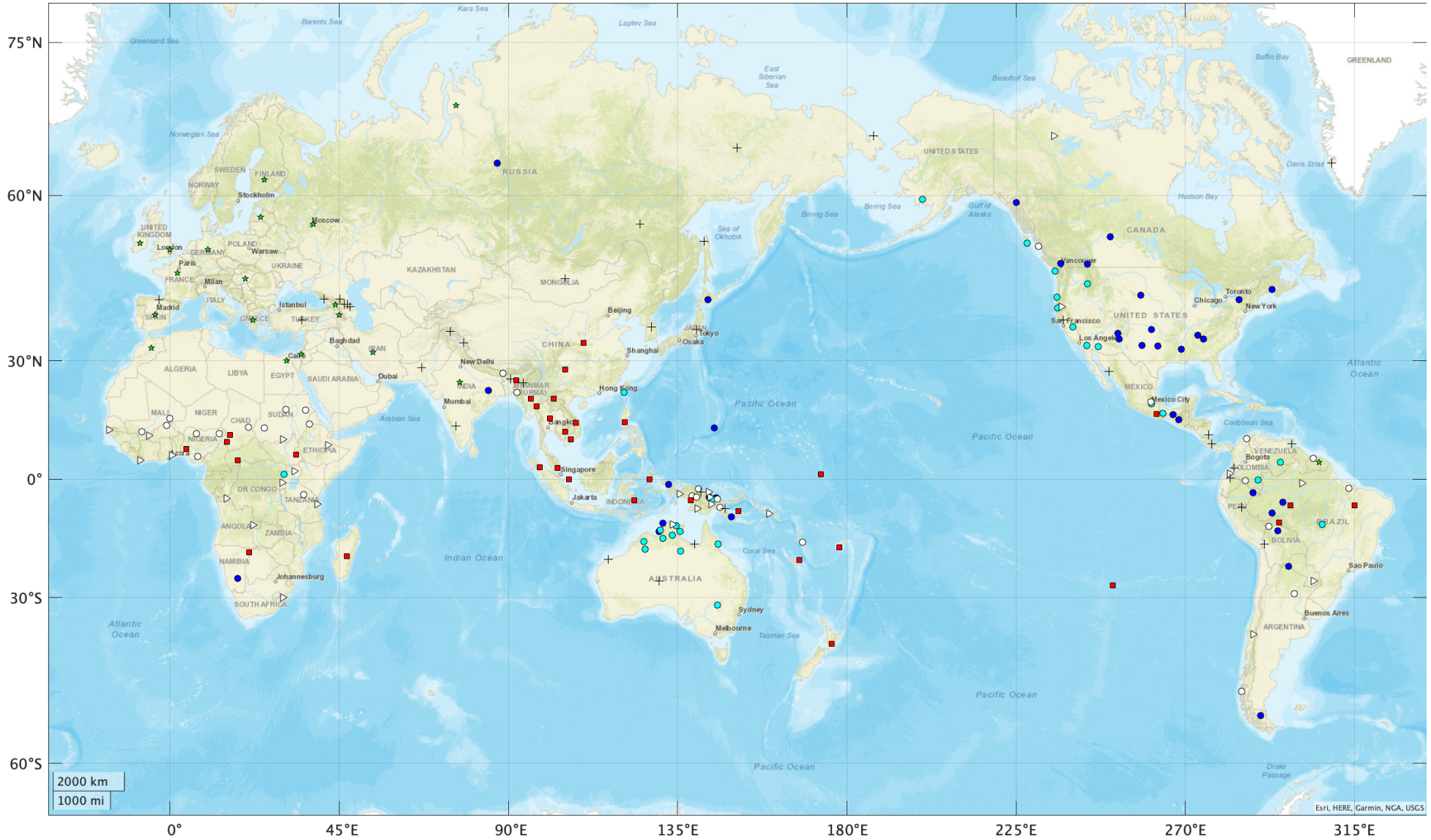

Figure 7: The classification results in C4 (i.e., from the 1st to the 145st dimension).

Blue circle: cluster 1; cyan circle: cluster 2; white circle: cluster 3; triangle: cluster 4; plus: cluster 5; star: cluster 6; rectangle: cluster 7 in C4 in Table 16.

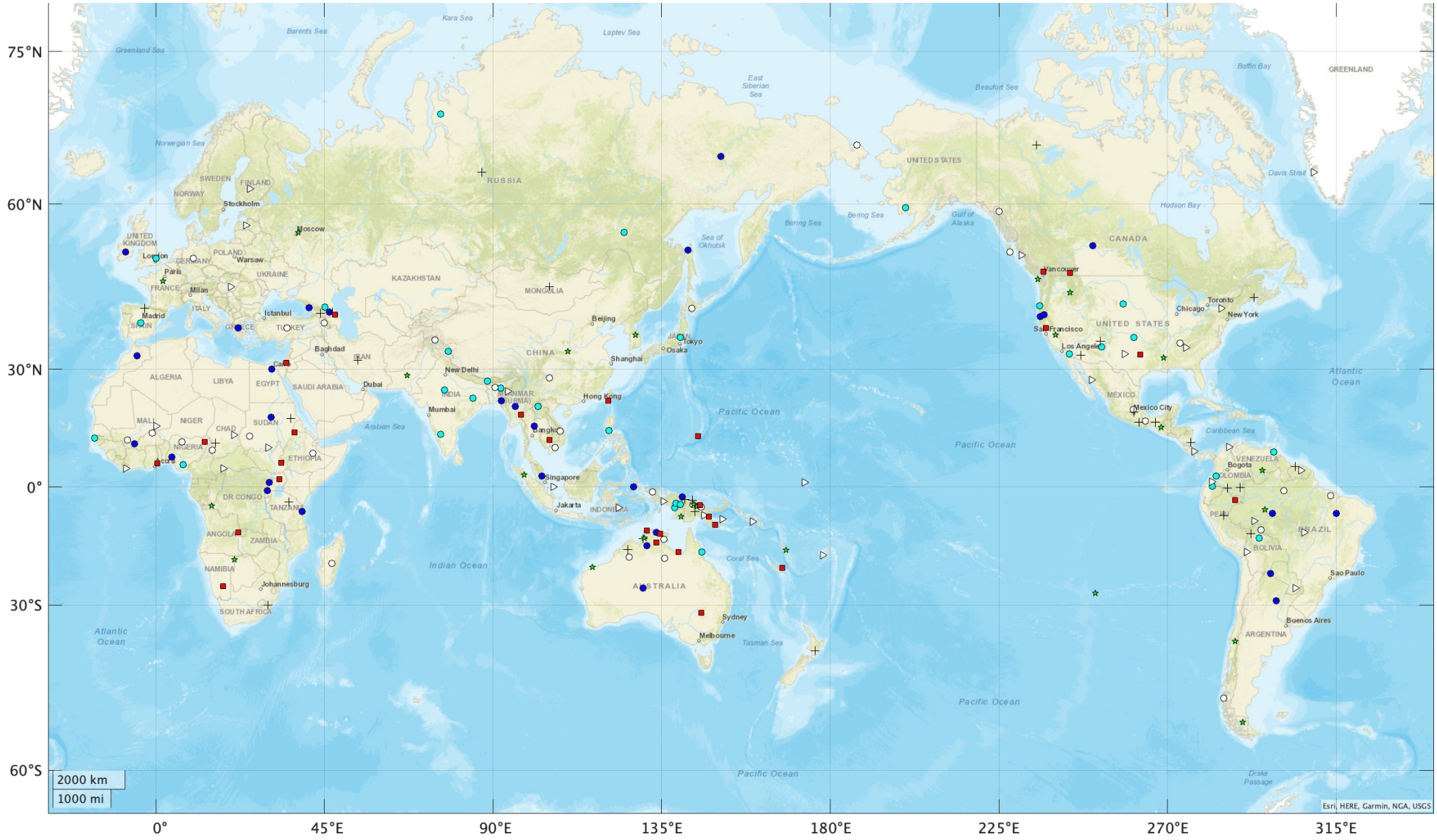

Figure 8: The classification results in C5 (i.e., from the 146st to the 202nd dimension).

Blue circle: cluster 1; cyan circle: cluster 6; white circle: cluster 3; triangle: cluster 4; plus: cluster 5; star: cluster 6; rectangle: cluster 7 in C5 in Table 16.

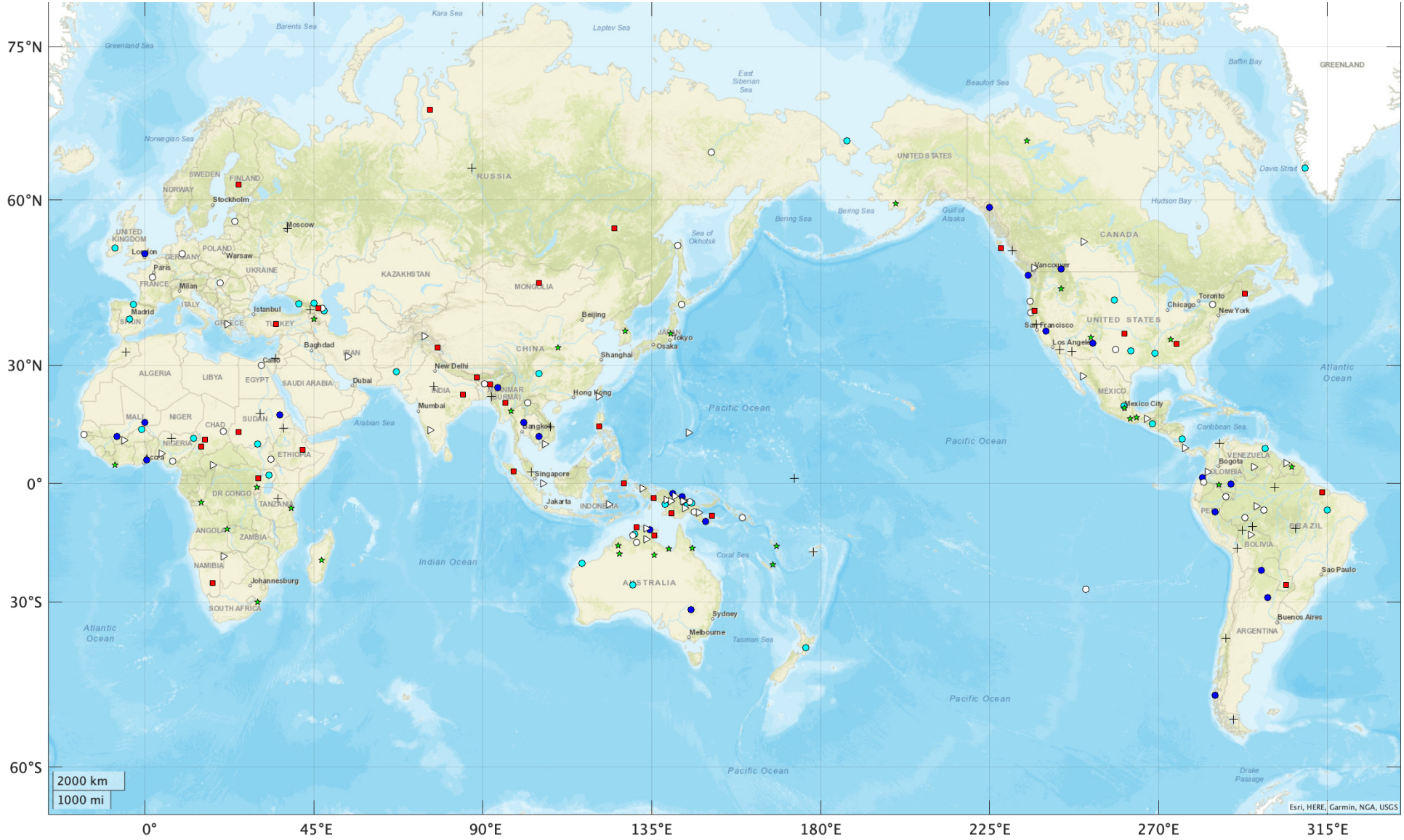

Figure 9: The classification results in C6 (i.e., from the 13st to the 202nd dimension).

Blue circle: cluster 1; cyan circle: cluster 2; white circle: cluster 3; triangle: cluster 4; plus: cluster 5; star: cluster 6; rectangle: cluster 7 in C6 in Table 16.

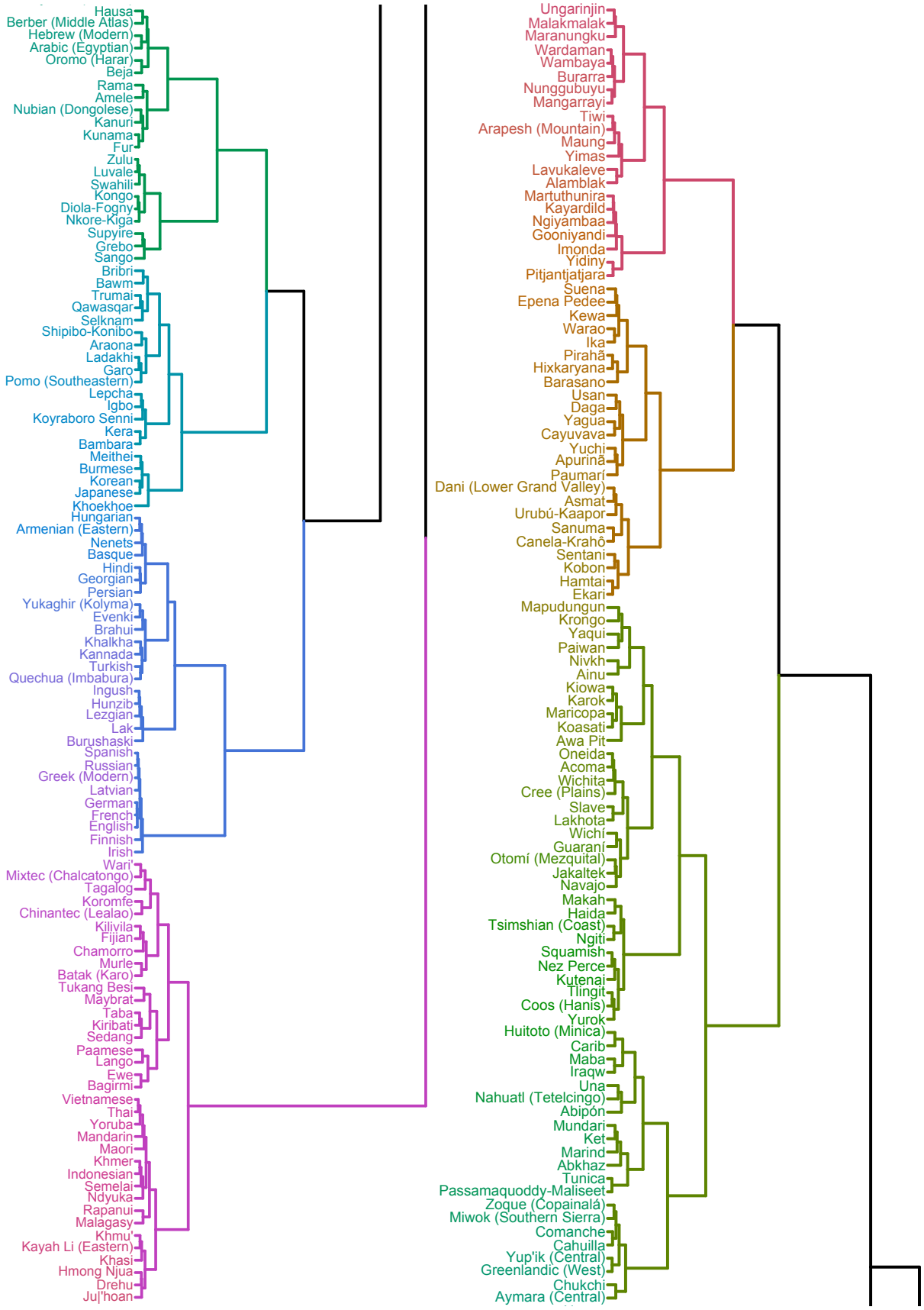

Figure 10: The hierarchical clustering result in C1 utilizing Ward method (Ward 1963).

Right: The upper part of the clustering result; Left: The lower part of clustering result.

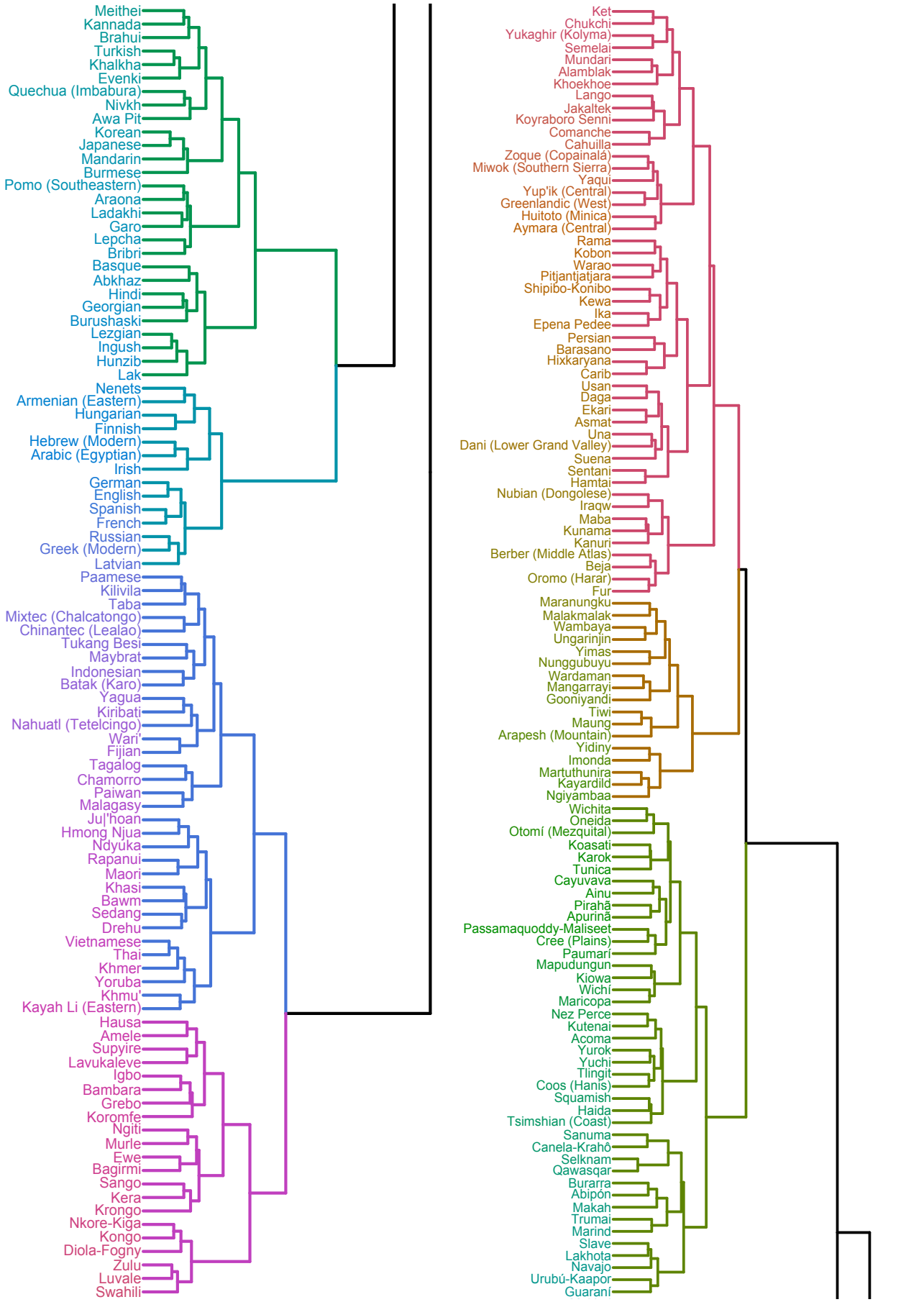

Figure11: The hierarchical clustering result in C2 utilizing Ward method (Ward 1963).

Right: The upper part of the clustering result; Left: The lower part of clustering result.

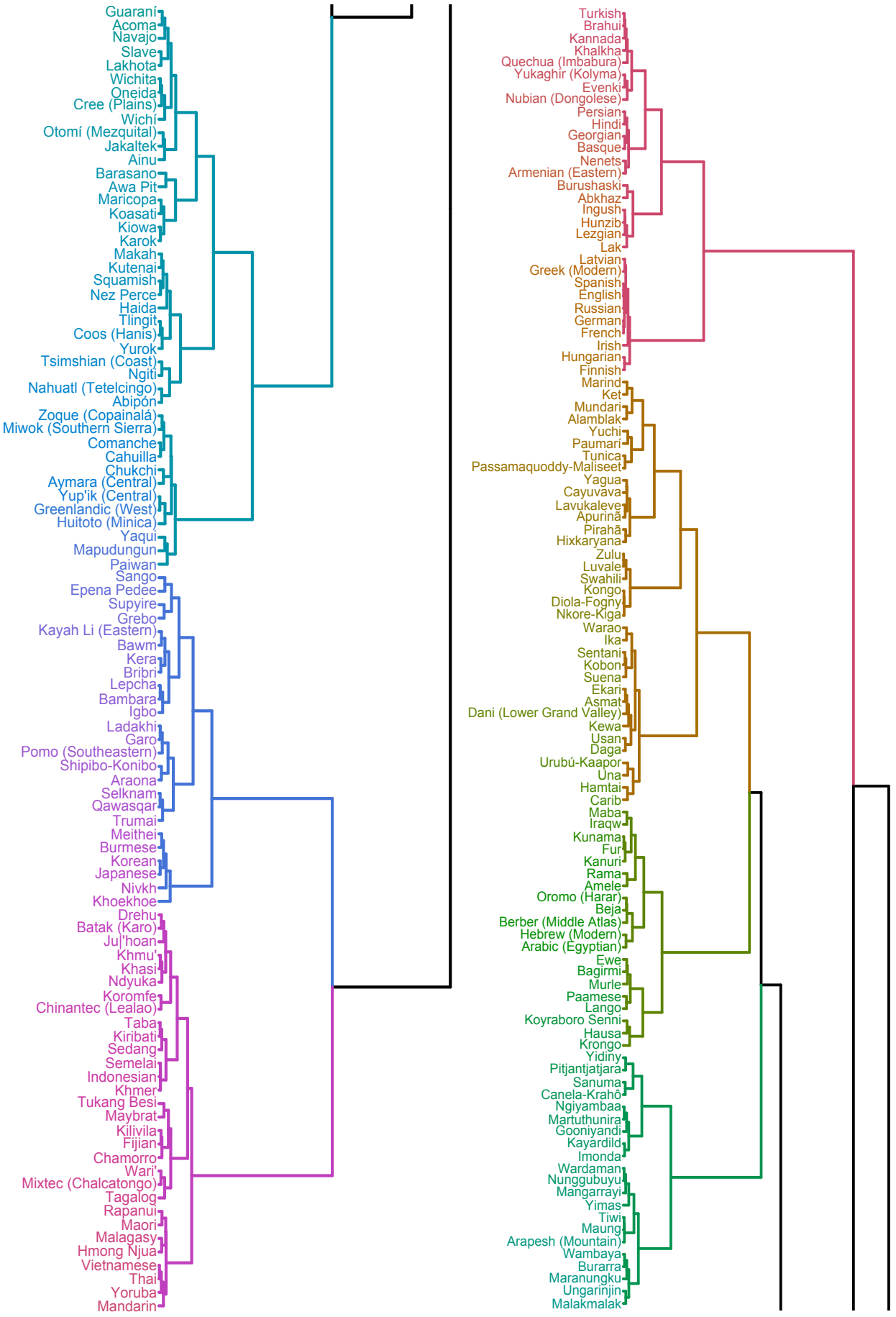

Figure12: The hierarchical clustering result in C3 utilizing Ward method (Ward 1963).

Right: The upper part of the clustering result; Left: The lower part of clustering result.

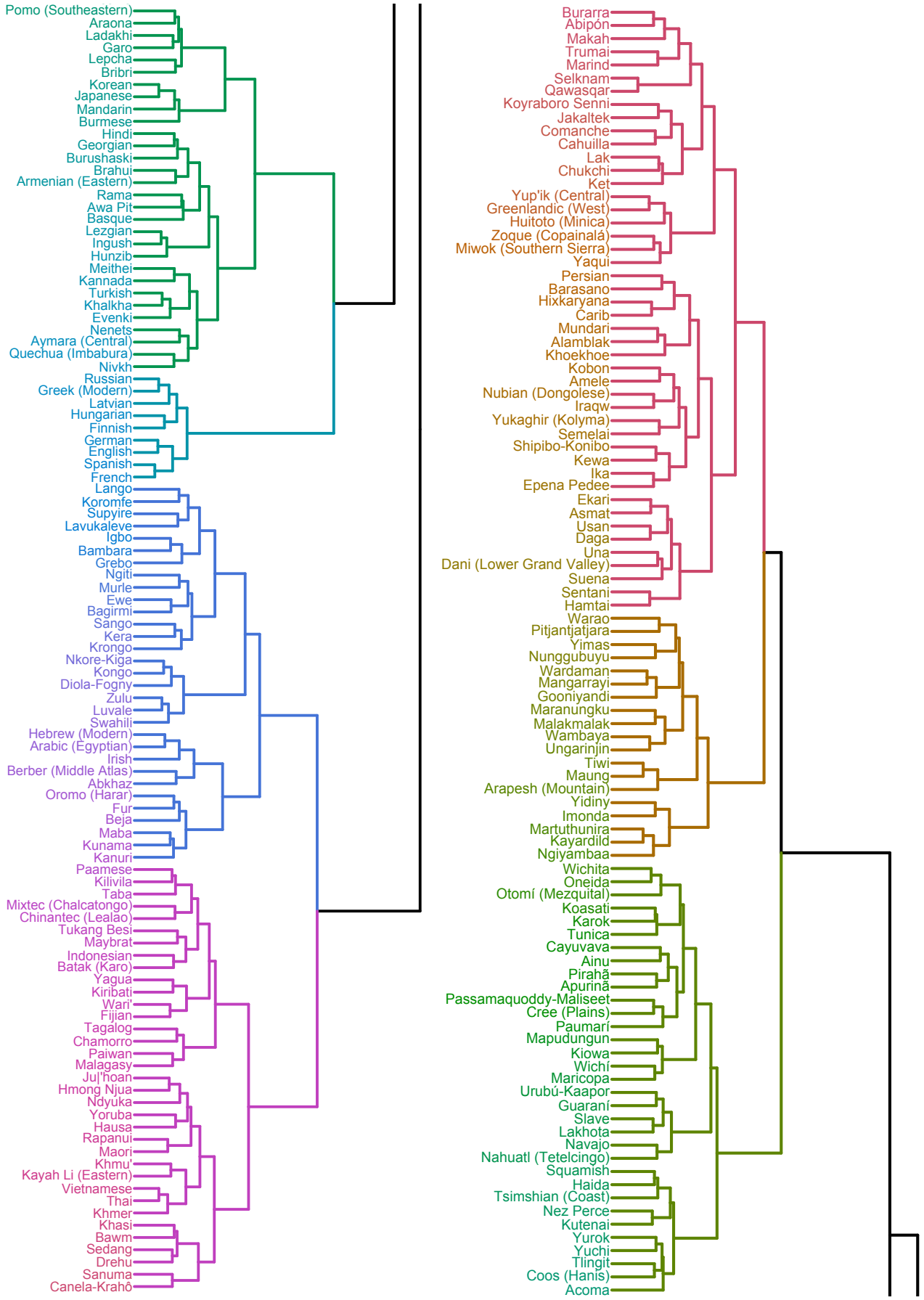

Figure 13: The hierarchical clustering result in C4 utilizing Ward method (Ward 1963).

Right: The upper part of the clustering result; Left: The lower part of clustering result.

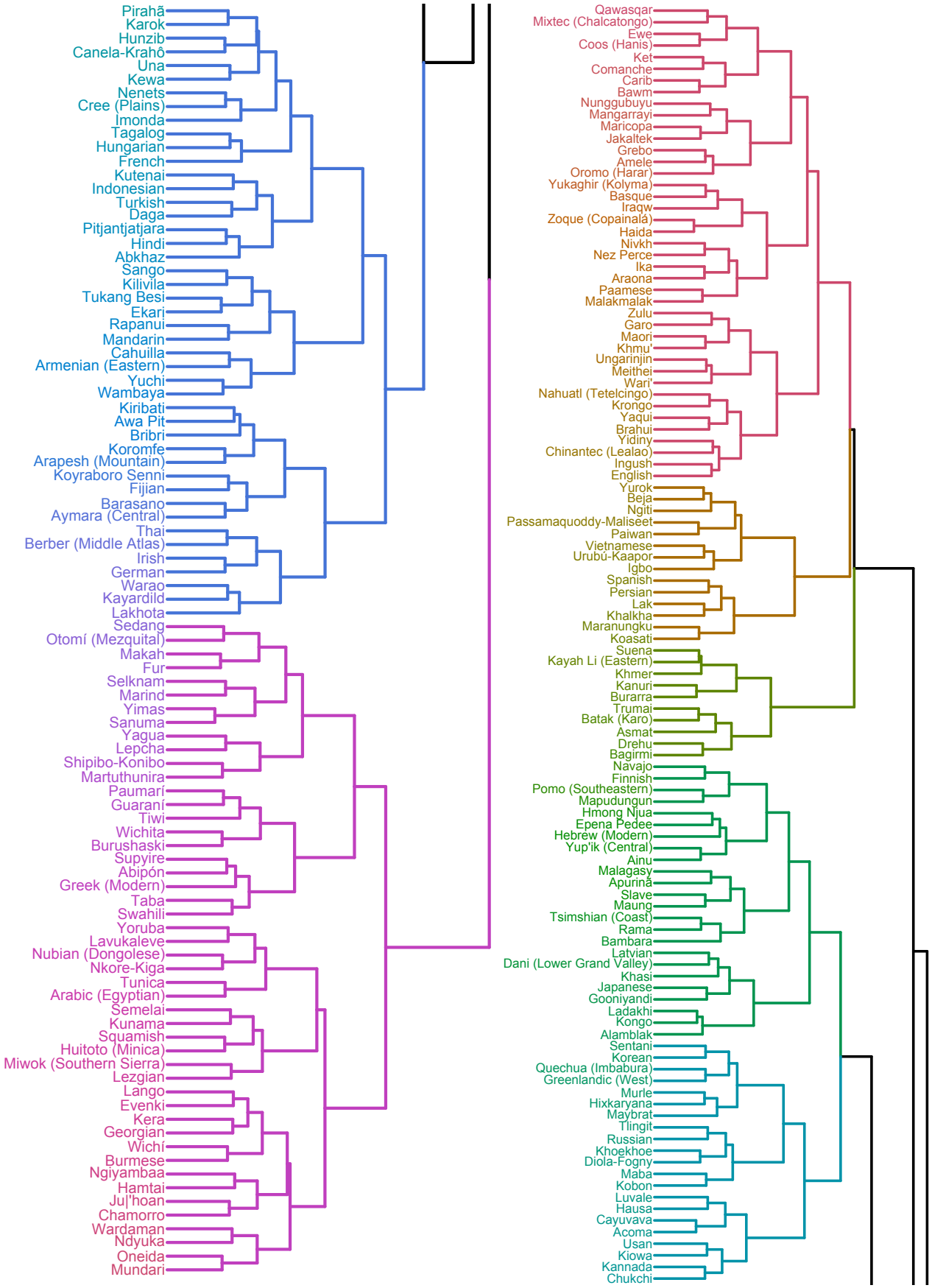

Figure 14: The hierarchical clustering result in C5 utilizing Ward method (Ward 1963).

Right: The upper part of the clustering result; Left: The lower part of clustering result.

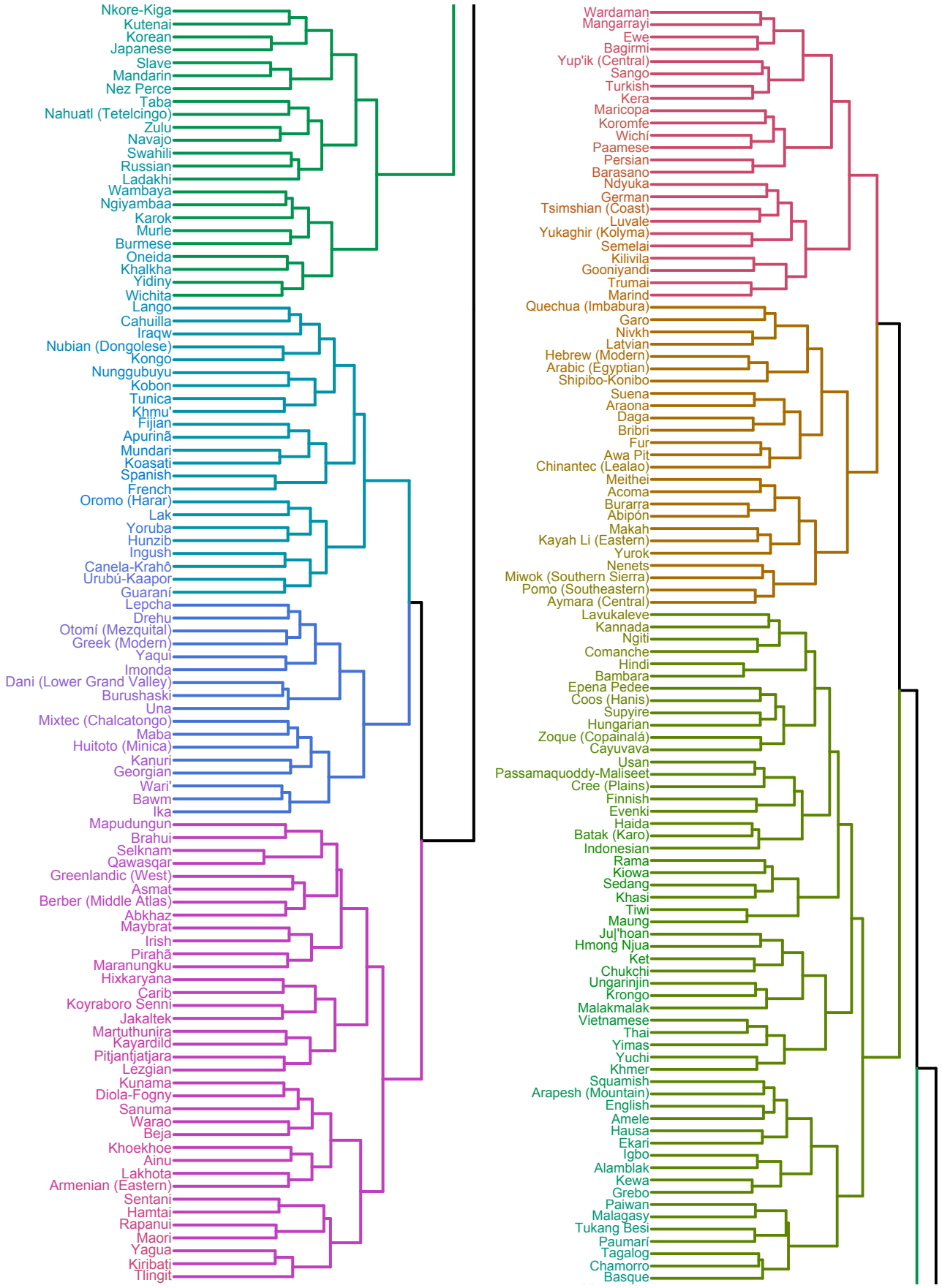

Figure 15: The hierarchical clustering result in C6 utilizing Ward method (Ward 1963).

Right: The upper part of the clustering result; Left: The lower part of clustering result.

Table 17. Features and 489 values in Data 1.

1A_1	14A_7	27A_3	39A_3	52A_3	69A_2	84A_1	98A_2	111A_4	126A_2	144R_8
1A_2	15A_3	28A_1	39A_5	53A_1	69A_4	84A_3	98A_4	112A_1	126A_3	144K_4
1A_3	15A_5	28A_2	40A_1	53A_4	69A_5	84A_6	99A_1	112A_2	127A_1	144B_3
1A_4	15A_7	28A_3	40A_2	53A_5	70A_1	85A_1	99A_2	112A_4	127A_2	144B_4
1A_5	15A_8	28A_4	40A_3	53A_6	70A_2	85A_2	99A_4	112A_6	127A_3	144D_2
2A_1	16A_1	29A_1	40A_5	53A_7	70A_4	85A_5	100A_1	113A_1	128A_1	144D_13
2A_2	16A_2	29A_2	41A_2	54A_1	70A_5	86A_1	100A_2	113A_2	128A_2	143E_1
2A_3	16A_4	29A_3	41A_3	54A_2	71A_1	86A_2	100A_4	113A_3	129A_1	143E_2
3A_1	16A_6	30A_1	42A_1	54A_4	71A_2	86A_3	100A_6	114A_1	129A_2	143E_4
3A_2	16A_7	30A_2	42A_2	55A_1	71A_3	87A_1	101A_1	114A_2	130A_1	143A_1
3A_3	17A_1	30A_3	42A_3	55A_2	71A_4	87A_2	101A_2	114A_3	130A_2	143A_2
3A_4	17A_4	30A_5	43A_1	55A_3	72A_1	87A_3	101A_4	114A_5	131A_1	143A_3
3A_5	17A_5	31A_1	43A_2	56A_1	72A_4	88A_1	101A_5	114A_7	131A_2	143A_4
4A_1	18A_1	31A_2	43A_3	56A_2	73A_1	88A_2	102A_1	115A_1	131A_3	143A_14
4A_2	18A_3	31A_3	43A_5	56A_3	73A_2	89A_1	102A_2	116A_1	131A_6	143A_15
4A_3	19A_1	32A_1	44A_1	57A_1	74A_1	89A_2	102A_5	116A_2	135A_1	144W_6
4A_4	19A_5	32A_2	44A_2	57A_2	74A_2	90A_1	103A_1	116A_6	136A_1	144Q_4
5A_1	20A_1	32A_3	44A_3	57A_4	74A_3	90A_2	103A_2	117A_1	136A_2	144L_3
5A_2	20A_7	33A_1	44A_6	58A_1	75A_1	90A_7	103A_4	117A_3	137A_1	144L_6
5A_3	21A_1	33A_2	45A_1	58A_2	75A_2	91A_1	103A_5	117A_4	138A_1	144L_15
5A_4	21A_5	33A_6	45A_2	59A_1	75A_3	91A_2	104A_1	117A_5	138A_2	144L_16
6A_1	22A_2	33A_7	46A_1	59A_2	76A_1	91A_3	104A_2	118A_1	138A_3	144H_4
6A_2	22A_3	33A_8	46A_2	59A_3	76A_2	92A_1	104A_3	118A_2	142A_1	143G_4
6A_4	22A_4	33A_9	47A_1	60A_4	76A_3	92A_2	104A_4	118A_3	142A_2	137B_1
7A_1	22A_5	34A_3	47A_2	60A_6	77A_1	92A_3	104A_5	119A_1	143F_1	137B_2
7A_2	23A_1	34A_4	48A_1	61A_2	77A_2	92A_6	105A_1	119A_2	143F_2	136B_1
7A_6	23A_2	34A_6	48A_2	62A_1	77A_3	93A_1	105A_2	120A_1	143F_4	136B_2
8A_1	23A_3	35A_3	48A_3	62A_2	78A_1	93A_2	105A_3	120A_2	90B_1	109B_1
8A_2	23A_4	35A_4	48A_4	62A_3	78A_2	94A_1	105A_4	121A_1	90C_1	109B_2
8A_4	24A_1	35A_5	49A_1	62A_7	78A_4	94A_2	106A_1	121A_2	144P_4	109B_4
9A_1	24A_2	35A_6	49A_2	62A_8	79A_1	94A_4	106A_2	121A_3	144J_7	109B_5
9A_2	24A_3	35A_7	49A_6	63A_1	79A_3	94A_5	106A_4	121A_4	144S_1	25B_2
9A_3	24A_4	35A_8	49A_7	63A_2	79A_4	95A_1	107A_1	122A_2	144S_2	21B_1
10A_1	25A_1	36A_1	49A_8	64A_1	80A_1	95A_4	107A_2	122A_4	144S_11	21B_2
10A_2	25A_2	36A_2	49A_9	64A_2	80A_3	95A_5	108A_1	123A_1	144X_4	108B_1
11A_1	25A_3	36A_3	50A_1	65A_1	81A_1	96A_1	108A_2	123A_2	144A_2	108B_4
12A_1	25A_5	36A_4	50A_2	65A_2	81A_2	96A_2	108A_3	123A_3	144A_7	58B_1
12A_2	26A_1	37A_1	50A_3	66A_1	81A_3	96A_4	109A_1	123A_4	144A_16	79B_2
12A_3	26A_2	37A_2	50A_4	66A_2	81A_7	96A_5	109A_3	124A_1	144A_18	79B_5
13A_1	26A_3	37A_4	50A_6	66A_4	82A_1	97A_1	109A_8	124A_2	144A_20	
13A_2	26A_4	37A_5	51A_1	67A_1	82A_2	97A_2	110A_1	124A_4	144A_21	
13A_3	26A_5	38A_1	51A_6	67A_2	82A_3	97A_3	110A_2	125A_1	144V_1	
14A_1	26A_6	38A_2	51A_9	68A_3	83A_1	97A_4	110A_3	125A_2	144I_1	
14A_2	27A_1	38A_4	52A_1	68A_4	83A_2	97A_5	111A_1	125A_3	144I_8	
14A_6	27A_2	38A_5	52A_2	69A_1	83A_3	98A_1	111A_2	126A_1	144R_1	

Note that the detail explanation can refer to Dryer and Haspelmath (2013a). The first symbols (e.g., 1A is “consonant inventory” in WALS [Maddieson 2013a]) correspond to the number of the feature in WALS, respectively, and each second symbols with underscore after the corresponding first symbol (e.g., _1 in 1A_1) is related to the value of the corresponding feature in WALS (e.g., _1 in 1A_1 is the value of “small” in [Maddieson 2013a]).

Table 18. Features and 391 values in Data 2.

1A_1	12A_1	23A_3	33A_1	42A_2	52A_3	65A_2	78A_1	105A_3	117A_3	130A_1
1A_2	12A_2	23A_4	33A_2	42A_3	53A_1	66A_1	78A_2	105A_4	117A_4	130A_2
1A_3	12A_3	24A_1	33A_6	43A_1	53A_4	66A_2	78A_4	106A_1	117A_5	131A_1
1A_4	13A_1	24A_2	33A_7	43A_2	53A_5	66A_4	79A_1	106A_2	118A_1	131A_2
1A_5	13A_2	24A_3	33A_8	43A_3	53A_6	67A_1	79A_3	106A_4	118A_2	131A_3
2A_1	13A_3	24A_4	33A_9	43A_5	53A_7	67A_2	79A_4	107A_1	118A_3	131A_6
2A_2	14A_1	25A_1	34A_3	44A_1	54A_1	68A_3	80A_1	107A_2	119A_1	135A_1
2A_3	14A_2	25A_2	34A_4	44A_2	54A_2	68A_4	80A_3	108A_1	119A_2	136A_1
3A_1	14A_6	25A_3	34A_6	44A_3	54A_4	69A_1	98A_1	108A_2	120A_1	136A_2
3A_2	14A_7	25A_5	35A_3	44A_6	55A_1	69A_2	98A_2	108A_3	120A_2	137A_1
3A_3	15A_3	26A_1	35A_4	45A_1	55A_2	69A_4	98A_4	109A_1	121A_1	138A_1
3A_4	15A_5	26A_2	35A_5	45A_2	55A_3	69A_5	99A_1	109A_3	121A_2	138A_2
3A_5	15A_7	26A_3	35A_6	46A_1	56A_1	70A_1	99A_2	109A_8	121A_3	138A_3
4A_1	15A_8	26A_4	35A_7	46A_2	56A_2	70A_2	99A_4	110A_1	121A_4	142A_1
4A_2	16A_1	26A_5	35A_8	47A_1	56A_3	70A_4	100A_1	110A_2	122A_2	142A_2
4A_3	16A_2	26A_6	36A_1	47A_2	57A_1	70A_5	100A_2	110A_3	122A_4	137B_1
4A_4	16A_4	27A_1	36A_2	48A_1	57A_2	71A_1	100A_4	111A_1	123A_1	137B_2
5A_1	16A_6	27A_2	36A_3	48A_2	57A_4	71A_2	100A_6	111A_2	123A_2	136B_1
5A_2	16A_7	27A_3	36A_4	48A_3	58A_1	71A_3	101A_1	111A_4	123A_3	136B_2
5A_3	17A_1	28A_1	37A_1	48A_4	58A_2	71A_4	101A_2	112A_1	123A_4	109B_1
5A_4	17A_4	28A_2	37A_2	49A_1	59A_1	72A_1	101A_4	112A_2	124A_1	109B_2
6A_1	17A_5	28A_3	37A_4	49A_2	59A_2	72A_4	101A_5	112A_4	124A_2	109B_4
6A_2	18A_1	28A_4	37A_5	49A_6	59A_3	73A_1	102A_1	112A_6	124A_4	109B_5
6A_4	18A_3	29A_1	38A_1	49A_7	60A_4	73A_2	102A_2	113A_1	125A_1	25B_2
7A_1	19A_1	29A_2	38A_2	49A_8	60A_6	74A_1	102A_5	113A_2	125A_2	21B_1
7A_2	19A_5	29A_3	38A_4	49A_9	61A_2	74A_2	103A_1	113A_3	125A_3	21B_2
7A_6	20A_1	30A_1	38A_5	50A_1	62A_1	74A_3	103A_2	114A_1	126A_1	108B_1
8A_1	20A_7	30A_2	39A_3	50A_2	62A_2	75A_1	103A_4	114A_2	126A_2	108B_4
8A_2	21A_1	30A_3	39A_5	50A_3	62A_3	75A_2	103A_5	114A_3	126A_3	58B_1
8A_4	21A_5	30A_5	40A_1	50A_4	62A_7	75A_3	104A_1	114A_5	127A_1	79B_2
9A_1	22A_2	31A_1	40A_2	50A_6	62A_8	76A_1	104A_2	114A_7	127A_2	79B_5
9A_2	22A_3	31A_2	40A_3	51A_1	63A_1	76A_2	104A_3	115A_1	127A_3	
9A_3	22A_4	31A_3	40A_5	51A_6	63A_2	76A_3	104A_4	116A_1	128A_1	
10A_1	22A_5	32A_1	41A_2	51A_9	64A_1	77A_1	104A_5	116A_2	128A_2	
10A_2	23A_1	32A_2	41A_3	52A_1	64A_2	77A_2	105A_1	116A_6	129A_1	
11A_1	23A_2	32A_3	42A_1	52A_2	65A_1	77A_3	105A_2	117A_1	129A_2	

Note that the detail explanation can refer to Dryer and Haspelmath (2013a). The first symbols (e.g., 1A is “consonant inventory” in WALS [Maddieson 2013a]) correspond to the number of the feature in WALS, respectively, and each second symbols with underscore after the corresponding first symbol (e.g., _1 in 1A_1) is related to the value of the corresponding feature in WALS (e.g., _1 in 1A_1 is the value of “small” in [Maddieson 2013a]).

Table 19. Features and 429 values in Data 3.

1A_1	13A_2	25A_3	35A_6	48A_1	59A_1	74A_1	97A_5	109A_3	122A_4	144V_1
1A_2	13A_3	25A_5	35A_7	48A_2	59A_2	74A_2	98A_1	109A_8	123A_1	144B_3
1A_3	14A_1	26A_1	35A_8	48A_3	59A_3	74A_3	98A_2	110A_1	123A_2	144B_4
1A_4	14A_2	26A_2	36A_1	48A_4	60A_4	75A_1	98A_4	110A_2	123A_3	143E_1
1A_5	14A_6	26A_3	36A_2	49A_1	60A_6	75A_2	99A_1	110A_3	123A_4	143E_2
2A_1	14A_7	26A_4	36A_3	49A_2	61A_2	75A_3	99A_2	111A_1	124A_1	143E_4
2A_2	15A_3	26A_5	36A_4	49A_6	62A_1	76A_1	99A_4	111A_2	124A_2	143A_1
2A_3	15A_5	26A_6	37A_1	49A_7	62A_2	76A_2	100A_1	111A_4	124A_4	143A_2
3A_1	15A_7	27A_1	37A_2	49A_8	62A_3	76A_3	100A_2	112A_1	125A_1	143A_3
3A_2	15A_8	27A_2	37A_4	49A_9	62A_7	77A_1	100A_4	112A_2	125A_2	143A_4
3A_3	16A_1	27A_3	37A_5	50A_1	62A_8	77A_2	100A_6	112A_4	125A_3	143A_14
3A_4	16A_2	28A_1	38A_1	50A_2	63A_1	77A_3	101A_1	112A_6	126A_1	143A_15
3A_5	16A_4	28A_2	38A_2	50A_3	63A_2	78A_1	101A_2	113A_1	126A_2	143G_4
4A_1	16A_6	28A_3	38A_4	50A_4	64A_1	78A_2	101A_4	113A_2	126A_3	137B_1
4A_2	16A_7	28A_4	38A_5	50A_6	64A_2	78A_4	101A_5	113A_3	127A_1	137B_2
4A_3	17A_1	29A_1	39A_3	51A_1	65A_1	79A_1	102A_1	114A_1	127A_2	136B_1
4A_4	17A_4	29A_2	39A_5	51A_6	65A_2	79A_3	102A_2	114A_2	127A_3	136B_2
5A_1	17A_5	29A_3	40A_1	51A_9	66A_1	79A_4	102A_5	114A_3	128A_1	109B_1
5A_2	18A_1	30A_1	40A_2	52A_1	66A_2	80A_1	103A_1	114A_5	128A_2	109B_2
5A_3	18A_3	30A_2	40A_3	52A_2	66A_4	80A_3	103A_2	114A_7	129A_1	109B_4
5A_4	19A_1	30A_3	40A_5	52A_3	67A_1	87A_1	103A_4	115A_1	129A_2	109B_5
6A_1	19A_5	30A_5	41A_2	53A_1	67A_2	87A_2	103A_5	116A_1	130A_1	25B_2
6A_2	20A_1	31A_1	41A_3	53A_4	68A_3	87A_3	104A_1	116A_2	130A_2	21B_1
6A_4	20A_7	31A_2	42A_1	53A_5	68A_4	91A_1	104A_2	116A_6	131A_1	21B_2
7A_1	21A_1	31A_3	42A_2	53A_6	69A_1	91A_2	104A_3	117A_1	131A_2	108B_1
7A_2	21A_5	32A_1	42A_3	53A_7	69A_2	91A_3	104A_4	117A_3	131A_3	108B_4
7A_6	22A_2	32A_2	43A_1	54A_1	69A_4	92A_1	104A_5	117A_4	131A_6	58B_1
8A_1	22A_3	32A_3	43A_2	54A_2	69A_5	92A_2	105A_1	117A_5	135A_1	79B_2
8A_2	22A_4	33A_1	43A_3	54A_4	70A_1	92A_3	105A_2	118A_1	136A_1	79B_5
8A_4	22A_5	33A_2	43A_5	55A_1	70A_2	92A_6	105A_3	118A_2	136A_2	
9A_1	23A_1	33A_6	44A_1	55A_2	70A_4	93A_1	105A_4	118A_3	137A_1	
9A_2	23A_2	33A_7	44A_2	55A_3	70A_5	93A_2	106A_1	119A_1	138A_1	
9A_3	23A_3	33A_8	44A_3	56A_1	71A_1	96A_1	106A_2	119A_2	138A_2	
10A_1	23A_4	33A_9	44A_6	56A_2	71A_2	96A_2	106A_4	120A_1	138A_3	
10A_2	24A_1	34A_3	45A_1	56A_3	71A_3	96A_4	107A_1	120A_2	142A_1	
11A_1	24A_2	34A_4	45A_2	57A_1	71A_4	96A_5	107A_2	121A_1	142A_2	
12A_1	24A_3	34A_6	46A_1	57A_2	72A_1	97A_1	108A_1	121A_2	143F_1	
12A_2	24A_4	35A_3	46A_2	57A_4	72A_4	97A_2	108A_2	121A_3	143F_2	
12A_3	25A_1	35A_4	47A_1	58A_1	73A_1	97A_3	108A_3	121A_4	143F_4	
13A_1	25A_2	35A_5	47A_2	58A_2	73A_2	97A_4	109A_1	122A_2	144X_4	

Note that the detail explanation can refer to Dryer and Haspelmath (2013a). The first symbols (e.g., 1A is “consonant inventory” in WALS [Maddieson 2013a]) correspond to the number of the feature in WALS, respectively, and each second symbols with underscore after the corresponding first symbol (e.g., _1 in 1A_1) is related to the value of the corresponding feature in WALS (e.g., _1 in 1A_1 is the value of “small” in [Maddieson 2013a]).

Table 20. Features and 439 values in Data 4.

1A_1	12A_3	24A_4	35A_3	44A_6	56A_1	70A_5	93A_1	105A_4	118A_3	135A_1
1A_2	13A_1	25A_1	35A_4	45A_1	56A_2	71A_1	93A_2	106A_1	119A_1	136A_1
1A_3	13A_2	25A_2	35A_5	45A_2	56A_3	71A_2	96A_1	106A_2	119A_2	136A_2
1A_4	13A_3	25A_3	35A_6	46A_1	57A_1	71A_3	96A_2	106A_4	120A_1	137A_1
1A_5	14A_1	25A_5	35A_7	46A_2	57A_2	71A_4	96A_4	107A_1	120A_2	138A_1
2A_1	14A_2	26A_1	35A_8	46A_4	57A_4	72A_1	96A_5	107A_2	121A_1	138A_2
2A_2	14A_6	26A_2	36A_1	47A_1	58A_1	72A_4	97A_1	108A_1	121A_2	138A_3
2A_3	14A_7	26A_3	36A_2	47A_2	58A_2	73A_1	97A_2	108A_2	121A_3	142A_1
3A_1	15A_3	26A_4	36A_3	48A_1	59A_1	73A_2	97A_3	108A_3	121A_4	142A_2
3A_2	15A_5	26A_5	36A_4	48A_2	59A_2	74A_1	97A_4	109A_1	122A_1	143F_1
3A_3	15A_7	26A_6	37A_1	48A_3	59A_3	74A_2	97A_5	109A_3	122A_2	143F_2
3A_4	15A_8	27A_1	37A_2	48A_4	60A_4	74A_3	98A_1	109A_8	122A_4	143F_4
3A_5	16A_1	27A_2	37A_3	49A_1	60A_6	75A_1	98A_2	110A_1	123A_1	144V_1
4A_1	16A_2	27A_3	37A_4	49A_2	61A_2	75A_2	98A_4	110A_2	123A_2	144X_4
4A_2	16A_4	28A_1	37A_5	49A_6	62A_1	75A_3	99A_1	110A_3	123A_3	144B_3
4A_3	16A_6	28A_2	38A_1	49A_7	62A_2	76A_1	99A_2	111A_1	123A_4	144B_4
4A_4	16A_7	28A_3	38A_2	49A_8	62A_3	76A_2	99A_4	111A_2	124A_1	143A_1
5A_1	17A_1	28A_4	38A_4	49A_9	62A_7	76A_3	100A_1	111A_4	124A_2	143A_2
5A_2	17A_4	29A_1	38A_5	50A_1	62A_8	77A_1	100A_2	112A_1	124A_4	143A_3
5A_3	17A_5	29A_2	39A_2	50A_2	63A_1	77A_2	100A_4	112A_2	125A_1	143A_4
5A_4	18A_1	29A_3	39A_3	50A_3	63A_2	77A_3	100A_6	112A_4	125A_2	143A_14
6A_1	18A_3	30A_1	39A_5	50A_4	64A_1	78A_1	101A_1	112A_6	125A_3	143A_15
6A_2	19A_1	30A_2	40A_1	50A_6	64A_2	78A_2	101A_2	113A_1	126A_1	143G_4
6A_4	19A_5	30A_3	40A_2	51A_1	65A_1	78A_3	101A_4	113A_2	126A_2	137B_1
7A_1	20A_1	30A_5	40A_3	51A_6	65A_2	78A_4	101A_5	113A_3	126A_3	137B_2
7A_2	20A_2	31A_1	40A_5	51A_9	66A_1	79A_1	102A_1	114A_1	127A_1	136B_1
7A_3	20A_7	31A_2	41A_2	52A_1	66A_2	79A_3	102A_2	114A_2	127A_2	136B_2
7A_6	21A_1	31A_3	41A_3	52A_2	66A_4	79A_4	102A_5	114A_3	127A_3	109B_1
8A_1	21A_5	32A_1	42A_1	52A_3	67A_1	80A_1	103A_1	114A_5	128A_1	109B_2
8A_2	22A_2	32A_2	42A_2	53A_1	67A_2	80A_3	103A_2	114A_7	128A_2	109B_4
8A_3	22A_3	32A_3	42A_3	53A_4	68A_2	87A_1	103A_4	115A_1	129A_1	109B_5
8A_4	22A_4	33A_1	43A_1	53A_5	68A_3	87A_2	103A_5	116A_1	129A_2	25B_2
9A_1	22A_5	33A_2	43A_2	53A_6	68A_4	87A_3	104A_1	116A_2	130A_1	21B_1
9A_2	23A_1	33A_6	43A_3	53A_7	69A_1	91A_1	104A_2	116A_6	130A_2	21B_2
9A_3	23A_2	33A_7	43A_4	54A_1	69A_2	91A_2	104A_3	117A_1	131A_1	108B_1
10A_1	23A_3	33A_8	43A_5	54A_2	69A_4	91A_3	104A_4	117A_3	131A_2	108B_4
10A_2	23A_4	33A_9	43A_6	54A_4	69A_5	92A_1	104A_5	117A_4	131A_3	58B_1
11A_1	24A_1	34A_3	44A_1	55A_1	70A_1	92A_2	105A_1	117A_5	131A_6	79B_2
12A_1	24A_2	34A_4	44A_2	55A_2	70A_2	92A_3	105A_2	118A_1	132A_7	79B_5
12A_2	24A_3	34A_6	44A_3	55A_3	70A_4	92A_6	105A_3	118A_2	134A_1	

Note that the detail explanation can refer to Dryer and Haspelmath (2013b). The first symbols (e.g., 1A is "consonant inventory" in WALS [Maddieson 2013b]) correspond to the number of the feature in WALS, respectively, and each second symbols with underscore after the corresponding first symbol (e.g., _1 in 1A_1) is related to the value of the corresponding feature in WALS (e.g., _1 in 1A_1 is the value of "small" in [Maddieson 2013b]).