HOKKAIDO UNIVERSITY

Title	Areality and Genealogy in Linguistic Typology: A Graphical Modeling Approach to Pacific Rim
Author(s)	ONO, Yohei
Citation	北方言語研究, 14, 125-154
Issue Date	2024-03-20
Doc URL	http://hdl.handle.net/2115/92088
Type	bulletin (article)
Additional Information	There are other files related to this item in HUSCAP. Check the above URL.
File Information	08_Ono.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Northern Language Studies 14: 125-154 (Japan Association of NoLS, 2024)

Areality and Genealogy in Linguistic Typology:

A Graphical Modeling Approach to Pacific Rim

Yohei ONO
(St. Luke's International University)

Keywords: Circum-Pacific, North Pacific Rim, language typology, inverse correlation, partial residuals

1. Introduction

Recent advances in the large-scale database, including World Atlas of Language Structures Online (Dryer and Haspelmath [eds.] 2013a, 2013b; abbreviated WALS hereafter) or AUTOTYP (Bickel et al. 2017), have promoted interdisciplinary research on linguistic typology by other researchers in information engineering or biology. Consequently, applying probability theory, including Bayesian modeling, aims to reveal areality and genealogy of languages globally that cannot be traced by lexicostatistics.

However, as demonstrated by Ono (2020), most of the "interdisciplinary" research was in vain because "missing values" in WALS or AUTOTYP are completely different from statistical definitions of missing values, which previous studies have led to imputing linguistically meaningless values on "missing values" in WALS or AUTOTYP.

The main objectives of this paper are to address unresolved problems in Ono (2020). The misuses in probability theory have been demonstrated in Ono (2020). Nonetheless, there are still issues regarding whether we can approach areality and genealogy of languages in the world by linguistic typology without imputing missing values, and if possible, what statistical analyses are appropriate and admissible from both linguistic and statistical points of view.

As an exploratory analysis, this paper will propose transformed format of original data and a similarity measure between languages and apply statistical analyses combining correspondence analysis, correlation analysis, graphical modeling, and clustering techniques, all of which are based on the linguistic requirements introduced in this paper.

From methodological viewpoints, we will show that graphical modeling approaches have desirable properties for exploring areal and genealogical information in the data on linguistic typology simultaneously, which will potentially contribute to areal linguistics from linguistic typology and visualizing such information in the humanities as a whole.

As a result, we have revealed Circum-Pacific clustering in WALS and similar clustering to Pacific Rim (Nichols 1994; Bickel and Nichols 2006), which suggests that statistical analyses are still promising for exploring areality and genealogy on languages in the world by linguistic typology and improving quality in WALS or AUTOTYP can potentially verify the hypotheses on North Pacific Rim (Miyaoka 1992) in future.

The remainder of this paper is organized as follows. Section 2 will overview previous studies and demonstrate why the statistical modeling approach is invalid from the viewpoint of both linguistics and statistics, focusing on “missing values” in linguistic typology. Moreover, we will consider alternatives to address areality and genealogy in languages using linguistic typology. Subsequently, problems will arise on how to measure some similarity between languages based on the datasets in linguistic typology containing numerous missing values, and we will propose transformed format of data and a similarity measure between languages that is admissible from the viewpoints of linguistic typology.

Section 3 will introduce basic information on Data 1, Data 2, Data 3, and Data 4 of WALS used in this paper, some of which have been utilized in previous studies.

Section 4 will introduce basic concepts to consider language as variable in Section 5, which will play a significant role in exploring areality and genealogy simultaneously.

Section 5 will develop new statistical analyses comprised of three linguistic requirements. We will discuss why three linguistic requirements are essential in areality and genealogy on linguistic typology and what statistical methods can fulfill three linguistic requirements, which results in validating our new statistical analyses from linguistic and statistical views. These discussions will lead us to statistical analyses combining correspondence analysis, correlation analysis, graphical modeling, and clustering techniques as a result. The evaluation criterion of clustering results is introduced in Section 6.

Section 7 will show classification results obtained by our proposed methods in Section 5. The results clearly support statistical analyses in previous studies (Ono and Whitman 2016; Whitman and Ono 2015) numerically, illustrate Circum-Pacific structure in linguistic typology, and detect similar clustering on Pacific Rim (Nichols 1994; Bickel and Nichols 2006), which suggests significance on improving the quality in database and developing statistical analyses appropriate for requirements in linguistic typology. Supplementary materials including detail results can be downloaded from https://researchmap.jp/ono_yohei/published_papers/44452272?lang=en.

Section 8 will discuss the significance of this paper from the viewpoints of both linguistics and statistics and how previous studies in the humanities can develop using statistical analyses proposed in this paper. Further directions of this paper are also discussed.

2. Background

Linguistic typology is still a promising key to revealing areality and genealogy of languages globally that cannot be captured by lexicostatistics because the scope of basic words is about one thousand years (Hymes 1960). Since the datasets are superficially easier to binarize in terms of the presence (1) or absence (0) of corresponding value on the feature in linguistic typology, it has attracted researchers in other fields, including computer science, information engineering, and biology, and numerous "interdisciplinary" studies have been conducted to numerically elucidate areality and genealogy applying statistical modeling including Bayesian statistical analysis (Daumé 2009; Daumé and Campbell 2007; Murawaki 2019; Takamura, Nagata, and Kawasaki 2016).

The background of "interdisciplinary" research is that most researchers are still attracted to similar success stories that applying Bayesian phylogenetic analysis has been overturning the establishment of evolution and divergence in biology (Drummond, Suchard, Xie, and Rambaut 2012). However, in the past two decades, the "interdisciplinary" studies have brought nothing substantively meaningful to linguistics because they have not examined assumptions underlying the datasets in linguistic typology, applied appropriate statistical methods, or developed statistical analyses corresponding to the assumptions in linguistic typology.

As demonstrated by Ono (2020), previous studies have misunderstood "missing values" in linguistic typology and abused statistical imputation methods to complement the "missing values," which have led their studies to meaningless statistical analyses as a result. The research on learning datasets in linguistic typology with probability distribution is paraphrased as the "statistical modeling approach" in the following section, and we will demonstrate why probability distribution is invalid in the "statistical modeling approach" on linguistic typology, using specific examples on northern languages in WALS datasets as explained in Ono (2020).

2.1 Why Statistical Modeling Approach is Invalid?

WALS data are comprised of features and their corresponding values. Table 1 shows four features and corresponding values of Ainu, Chukchi, Khalkha, and Navajo in WALS, respectively. Each feature is as follows: 12A: Syllable structures (Maddieson 2013); 14A: Fixed stress locations (Goedemans and van der Hulst 2013); 30A: Number of genders (Corbett 2013); and 47A: Intensifiers and reflexive pronouns (König, Siemund, and Töpper 2013). Furthermore, 12A consists of following values: 1: Simple syllable structure; 2: Moderately complex syllable structure; 3: Complex syllable structure, 14A consists of following values: 1: No fixed stress (mostly weight-sensitive stress); 2: Initial stress; 3: Second; 4: Third; 5: Antepenultimate; 6: Penultimate; 7: Ultimate stress, 30A consists of following values: 1: None; 2: Two; 3: Three; 4: Four; 5: Five or more, and 47A consists of following values: 1: Intensifiers and reflexive pronouns are formally identical; 2: Intensifiers and reflexive pronouns are formally differentiated.

In Table 1, NA represents originally blank data in WALS that comprise more than 70 percent of the datasets (Murawaki 2019: 202). As Ono (2020) demonstrated, the "statistical modeling approach" did not deal with NA in the linguistic context of WALS but distorted NA into the statistical context, which led to meaningless statistical analyses as a result.

In linguistic context, NA or blank data in WALS represents that the language possesses some linguistic characteristics that the features in WALS cannot precisely describe. But, in the context of statistics, NA conforms to represent that the language possesses some linguistic characteristics that the features in WALS can describe, but researchers have not yet observed for some reason (e.g., lack of descriptive research on the language or some environment around the language).

Table 1. Example of WALS data. Each feature and the corresponding missing values in Ainu, Chukchi, Khalkha, and Navajo cited from Ono (2020: 63).

Name	12A	14A	30A	47A
Ainu	2 Moderately complex	NA	1 None	NA
Chukchi	3 Complex	2 Initial	1 None	NA
Khalkha	NA	1 No fixed stress	1 None	1 Identical
Navajo	2 Moderately complex	1 No fixed stress	NA	2 Differentiated

Table 2. Result of statistical imputation to Table 1 in Ainu, Chukchi, Khalkha, and Navajo cited from Ono (2020: 75).

Name	12A	14A	30A	47A
Ainu	2 Moderately complex	1 No fixed stress	1 None	2 Differentiated
Chukchi	3 Complex	2 Initial	1 None	2 Differentiated
Khalkha	2 Moderately complex	1 No fixed stress	1 None	1 Identical
Navajo	2 Moderately complex	1 No fixed stress	1 None	2 Differentiated

For example, Maddieson (2013) explains 12A: Syllable structures on the combination of strings for consonant and vowel sound symbols as the criterion for complexity, and states, “Languages which permit a single consonant after the vowel and/or allow two consonants to occur before the vowel, but obey a limitation to only the common two-consonant patterns described above, are counted as having moderately complex syllable structure.” Thus, Ainu (Simeon 1969: 754) and Navajo (Sapir and Hoijer 1967: 3) correspond to 2: Moderately complex syllable structure.

However, Svantesson (2003: 158) states on Khalkha, “The maximal syllable structure is CVVCCC, i.e., the vowel kernel may be preceded by at most one consonant and followed by a cluster of up to three consonants. The vowel can be short, long, or diphthong. In non-initial syllables, it can also be a non-phonemic schwa vowel. Onsetless syllables occur only word-initially. Whether a consonant combination can form a syllable coda or not depends on the phonetic properties of the consonants. Permitted types of coda include voiced + voiceless consonant, e.g., *daws* [taws] ‘salt,’ *alt* [aɮ^{h}t] ‘gold,’ *bügd* [pugt] ‘all;’ nasal + stop or affricate, e.g., *xünd* [xunt] ‘heavy,’ *möngg* [moŋg] ‘silver,’ *myanggh* [m^{j}aŋɢ] ‘thousand;’ fricative + stop or affricate, e.g., *tsast* [ts^{h}as^{h}t] ‘snowy.’ Three-consonant codas consist of a voiced consonant followed by a fricative + stop or affricate, e.g., *ilst* [iɮs^{h}t] ‘sandy.’”

Thus, the syllable structure of Khalkha cannot be precisely captured by the string combination for consonant and vowel sound symbols but should be analyzed from the phonetic properties of the consonants restricting possible syllable structures. Consequently, the value in Khalkha is marked as NA in 12A in Table 1.

Thus, NA of 12A on Khalkha represents that the language possesses some linguistic characteristics that the features in WALS cannot precisely describe.

However, in the context of statistics, NA of 12A on Khalkha conforms to represent that the language possesses some linguistic characteristics that the features in WALS can describe, but researchers have not yet observed for some reason; in other words, “statistical modeling approach” will assume that NA of 12A on Khalkha is predetermined as 1: Simple syllable structure or 2: Moderately complex syllable structure or 3: Complex syllable structure but linguistic researchers have not yet observed on the specific value of syllable structure for Khalkha.

As Ono (2020) demonstrated, the “statistical modeling approach” will require imputing the original NA or blank data in WALS by other values; 1: Simple syllable structure, 2: Moderately complex syllable structure, 3: Complex syllable structure in our case. Table 2 is obtained by imputing NA in Table 1 by statistical imputation methods used in previous studies (Murawaki 2019). This resulted in replacing the original NA or blank data on 12A in Khalkha by 2: Moderately complex syllable structure.

This result violates the original NA or blank data on 12A in Khalkha in the linguistic context. Again, the syllable structure on Khalkha should be understood in terms of the underlying phonetic properties of the consonants, which the string combination for consonant and vowel sound in Maddieson (2013) cannot precisely capture and is marked as NA on 12A in Khalkha. The analysis can apply to other NAs in Table 1 (Ono 2020: 63–69). In this subsection, we have demonstrated that the “statistical modeling approach” is invalid because the original NA or blank data in WALS, different from NA in the statistical context, are imputed by other existing values in the feature, which cannot be supported from the substantive viewpoints of linguistics.

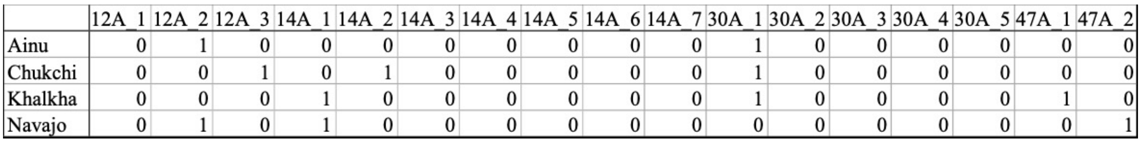
However, problems will soon arise as to whether statistical analyses can approach areality and genealogy on languages by linguistic typology without imputing NA and, if possible, what statistical methods are appropriate or admissible from both linguistic and statistical points of view.

Primarily, we need to address how to measure similarity between languages in WALS datasets without probability imputation, in other words, with a more descriptive way of statistics.

2.2 How to Transform Data and Measure Similarity Between Languages?

The previous subsection showed that replacing or imputing "NA" with other values in the feature was inappropriate from the viewpoints of both linguistics and statistics. Our next question is how we should measure the degree of similarity between languages in linguistic typology.

Table 3. Transformed data of Table 1. Presence of each value is coded as 1 and absence as 0.



	12A_1	12A_2	12A_3	14A_1	14A_2	14A_3	14A_4	14A_5	14A_6	14A_7	30A_1	30A_2	30A_3	30A_4	30A_5	47A_1	47A_2
Ainu	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Chukchi	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0
Khalkha	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
Navajo	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1

We will propose to transform Table 1 as Table 3 where the presence of each value is coded as 1 and absence as 0. For example, in 47A (Fixed stress locations), Ainu and Chukchi show NA, respectively, because both Ainu and Chukchi should be understood in terms of the affix, not the pronoun (Ono 2020: 67–69). However, the main concern in feature 47A is whether intensifiers and reflexive pronouns are formally identical or differentiated in the languages.

From the viewpoint of logic, the languages that represent intensifiers and reflexive by a concept other than pronoun (e.g., affix) are out of scope in 47A. Thus, Table 3 transforms Table 1 into each value in terms of presence and absence, taking into account the logical scope of the corresponding value.

Transforming our original data into the format shown in Table 3 is useful as follows: For example, in 14A (Fixed stress locations), Ainu and Japanese show NA, respectively, because both Ainu and Japanese should be understood in terms of the pitch accent system, not the stress accent system. If we code the presence of "NA" in 14 A, we could consider that Ainu and Japanese are similar in 14A because both languages show NA. However, these are irrelevant in the stress system (i.e., the scope of each value in 14 A), which leads to irrational measurement between languages in linguistic typology. Incorporating the idea of scope in logic, we can here exclude some possibilities considering languages to be similar to each other as they cannot be analyzed from certain linguistic features.

Thus, the intuitive similarity measure that we propose is 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, there are some biases in this measure that needs to be revised from the viewpoints of statistics as explained in Section 5.1

3. Materials

In this Section, we will introduce the WALS datasets used in this paper in detail. Previous studies (Ono et al. 2017, 2018; Ono and Whitman 2016; Whitman and Ono 2015, 2017) have suggested that word-order features are highly correlated in the datasets of linguistic typology, which can be classified into "prepositional-type," "postpositional-type," and "adpositionless-type" as a result. Furthermore, as exploratory analyses, Whitman and Ono (2015) and Ono and Whitman (2016) have applied statistical classifications to WALS data (Dryer and Haspelmath [eds.] 2013a), excluding all word-order features except those related to Adjective and Negation because these features on Adjective and Negation are less correlated in word-order features. As a result, Whitman and Ono (2015) and Ono and Whitman (2016) have found some areal and genetic classifications on languages in WALS, which suggested that most word-order features can mask areality and genealogy in linguistic typology.

The World Atlas of Language Structures Online (WALS) 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. Note that interested reader can refer in supplemental materials.

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 from Data 1 as "Data 2", those removing part of word-order parameters, which Whitman and Ono (2017) identified as the main component, as "Data 3", which included the parameters on adjective and negation that Dryer (1992) indicated, retained[1]. Data 2 contains 391 values and 201 languages, and Data 3 has 429 values and 201 languages[2].

Furthermore, Data 4 is comprised of 203 languages in WALS data updated (Dryer and Haspelmath [eds.] 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. Note that we have described the detail explanation about our data in supplemental materials.

4. Concepts

As indicated in the previous section, Whitman and Ono (2015) and Ono and Whitman (2016) have revealed some areality and genealogy in linguistic typology using WALS datasets. They have applied hierarchical clustering analyses (Ward 1963) to correlation

[1] 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.

[2] Data 2 and Data 3 are the same data used in Whitman and Ono (2015) and Ono and Whitman (2016).

similarity obtained by correspondence analysis to the WALS dataset explained in the previous section. However, Whitman and Ono (2015) and Ono and Whitman (2016) cannot explain why their statistical analyses are valid or what statistical methods can improve areality and genealogy in linguistic typology from the theoretical background of statistics, to which we will show some answers in this paper.

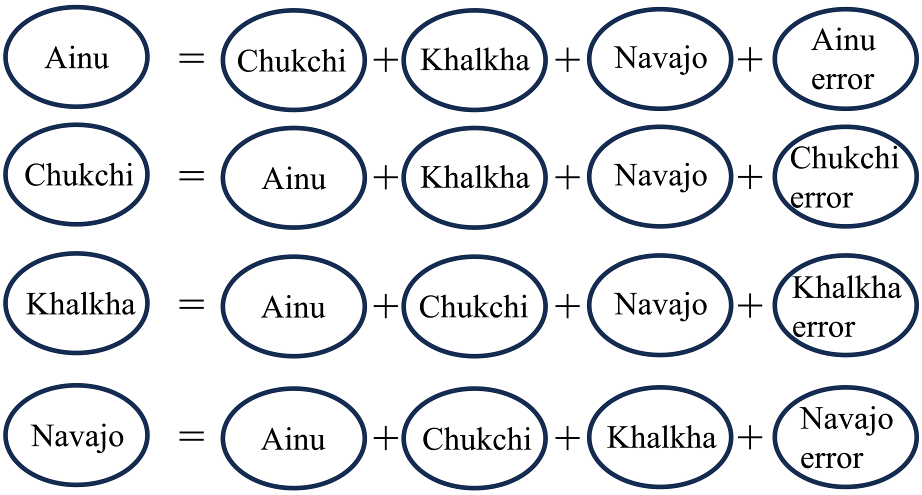

Ainu = Chukchi + Khalkha + Navajo + Ainu error
Chukchi = Ainu + Khalkha + Navajo + Chukchi error
Khalkha = Ainu + Chukchi + Navajo + Khalkha error
Navajo = Ainu + Chukchi + Khalkha + Navajo error

Figure 1. The basic image of the relationship between one language and the other languages in linguistic typology. For example, Ainu corresponds to linguistic typological characteristics in the Ainu language, and Ainu error corresponds to linguistic typological characteristics in the Ainu language that cannot be explained by linguistic typological characteristics in the other languages (i.e., Chukchi, Khalkha, and Navajo in this case).

Thus, this section will introduce some basic concepts for clarifying why statistical analyses in the previous studies are valid for areality and genealogy in linguistic typology and what part of our statistical analyses can improve Whitman and Ono (2015) and Ono and Whitman (2016). Section 2 allowed us to calculate the similarity between languages in linguistic typology data. In the first step, we will consider how the similarities should be analyzed from the viewpoints of both linguistics and statistics.

Considering areality and genealogy of languages in linguistic typology, it is trivial from a linguistic point of view that languages influence each other. If we dare to illustrate with statistical concepts, our situations can be pictured in Figure 1, where "linguistic typological characteristics in language A" is just abbreviated as "language A," and "linguistic typological characteristics in language A that cannot be explained by linguistic typological characteristics in the other languages" as "language A error."[3]

[3] Using the notation of "function," Figure 1 can be represented as follows: Ainu = f(Chukchi, Khalkha, Navajo) + Ainu error; Chukchi = f(Ainu, Khalkha, Navajo) + Chukchi error; Khalkha = f(Ainu, Chukchi, Navajo) + Khalkha error; Navajo = f(Ainu, Chukchi, Khalkha) + Navajo error. As illustrated in the following section, "function" statistically corresponds to the result of multiple linear regression,

Thus, language can be viewed not only as a subject but also as a variable that affects other variables (i.e., languages in our case), which has been absent in previous studies in statistical typology. As explained in the following section, this basic concept will enable us to decompose "linguistic typological characteristics in language A" (i.e., "language A") into two parts and their correlations: "linguistic typological characteristics in language A that can be explained by linguistic typological characteristics in the other languages" and "linguistic typological characteristics in language A that cannot be explained by linguistic typological characteristics in the other languages" (i.e., "language A error"). The latter will correspond to linguistic typological characteristics explained not by linguistics but by a factor other than linguistics (e.g., geographical factor).

The following section will demonstrate how beneficially this basic concept works to analyze areality and genealogy on WALS data in exploratory ways.

5. Basic Framework in Statistical Analyses

In this section, we will illustrate a basic framework for how to address areality and genealogy in linguistic typology based on three linguistic requirements for statistical analyses.

First, we will discuss that the similarity measure on a pair of languages should transform into numerical numbers or vectors on each language, whereby quantification results are required as unbiased by heterogeneous total numbers of languages or values included in each value or language, respectively. Thus, the first requirement can be called "correction on frequencies." We will conclude that correspondence analysis (Benzécri et coll. 1973) is the admissible statistical method that satisfies correction on frequencies under the assumption that our datasets are sufficiently representative of languages in the world from viewpoints of linguistic typology.

Second, we will discuss how to group or cluster the languages based on results calculated by correspondence analysis, whereby analysis of profile will be required as the similarity between languages in linguistic typology explicitly or implicitly rather than Euclidean distance. Thus, the second requirement can be called "profile view." We will conclude that correlation analysis is more appropriate for "profile view" rather than "distance view," including Euclidean distance.

Third, we will discuss how to address areality and genealogy in linguistic typology using correlation similarity obtained from correspondence analysis, whereby language will not only be a subject but also a variable for the other languages in linguistic typology in a more statistical sense. Thus, the third requirement can be called "language as variable." We will conclude that graphical modeling is more appropriate for "language as variable" because graphical modeling can analyze both areality and genealogy in linguistic

and error corresponds to residual in each multiple linear regression.

typology for exploratory purposes, especially the former of which will be an inverse correlation in our data from our discussions.

These three linguistic requirements will lead us to statistical analyses combining correspondence analysis, correlation analysis, graphical modeling, and clustering techniques as a result.

5.1 Linguistic Requirement (1): Correspondence Analysis

Section 2 illustrated how to measure similarity between pairs of language in WALS. Thus, we assume to have similarity data on a pair of languages with the number of rows and columns equal to the number of languages. However, Figure 1 required us transforming the similarity between pairs of languages into numerical numbers or vectors in each language, which is the focus of this subsection.

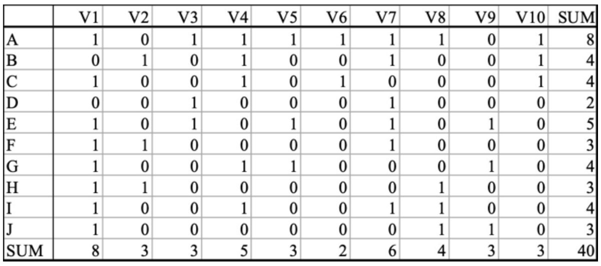
Following our proposed data in Section 2.2, one intuitive similarity measure can be explained as follows: we count agreement of the corresponding values as 1 or disagreement of the corresponding values as 0 in pairs of languages or values. However, as explained in this subsection, the intuitive similarity measure needs to be revised regarding some biases in the similarity data, which corresponds to the first requirement, "correction on frequencies" in this 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.

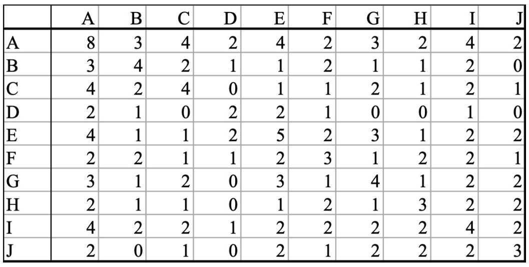
For example, Table 4 shows tentative example of data between languages and values, and calculated similarity between languages. In left side of Table 4, A language contains the largest number of values in all languages (i.e., 8 values), which resulting in A language more similar to another language in right side of Table 4. Furthermore, I language contains smaller numbers of values (i.e., 4 values) but they are more "popular" value (i.e., V1, V4, V7, and V8), which resulting in I language more similar to another language.

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 in details in supplementary materials, chi statistic is one promising solution to "correction on frequencies." Since supplementary materials will provide detail discussions about chi statistics, we will not deal with more specific content about chi statistic here.

Table 4. Left: tentative example of data between languages (row) and values (column). Right: calculated similarity between languages. Presence of value is coded as 1 and absence of value as 0.



	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



	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

Notably, chi statistic can be considered as measure of strength between languages and values, and 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.

Since the language can be considered as consisting of multidimensional factors, the obtained vector of the language can visualize multidimensional disposition in space, especially Euclidean space in the case of correspondence analysis, which will be essential in our second requirement of “profile view” in next subsection. Thus, correspondence analysis is an admissible statistical method for our exploratory analysis if we can assume that our data can be considered as sufficient representative of languages globally.

By applying correspondence analysis to our proposed data in Section 2.2, two statistics can be calculated: the coordinates $x_{ik}, i = 1, 2, \ldots, n; k = 1, 2, \ldots, n - 1$, where i corresponds to each language and the coordinates $y_{jk}, j = 1, 2, \ldots, p; k = 1, 2, \ldots, n - 1$, where i corresponds to each language, j corresponds to each value, k corresponds to dimension obtained by correspondence analysis, n is the number of languages and p is the number of values in our data (i.e., n is 201 in Data 1-3, 203 in Data 4, and p is 489 in Data 1, 391 in Data 2, 429 in Data 3, and 439 in Data 4), respectively; the eigenvalue in each dimension λ_k that is applied as $\omega_k = \lambda_k / \sum_{m=1}^{n-1} \lambda_m$ in following analyses.

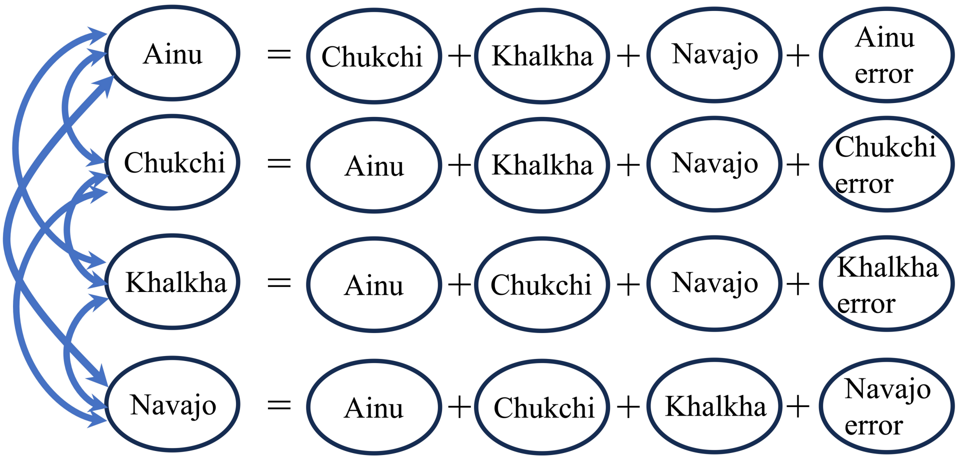
5.2 Linguistic Requirement (2): Correlation Analysis

As shown in the previous subsection, correspondence analysis can transform similarity information between languages in WALS into numerical vectors on each language that can be admissible if our data are sufficiently representative of languages and features in WALS. Thus, this subsection will assume numerical vectors in each language that correspond to Ainu, Chukchi, Khalkha, and Navajo in Figure 2 as an example[4]. Thus, our

[4] $\sqrt{\omega_k} x_{ik}$ corresponds to the numerical vectors and $i = 1$ (Ainu), $i = 2$ (Chukchi), $i = 3$ (Khalkha), and $i = 4$ (Navajo) for example.

next question is how to measure some similarity or dissimilarity between languages (i.e., blue bidirectional arrows in Figure 2) based on the vectors calculated by correspondence analysis.

Figure 3 is tentatively results, in which the first dimension (i.e., $k = 1$ in our case) corresponds to the quantified degree of "head-marking," the second dimension (i.e., $k = 2$ in our case) to "postpositional," and the third dimension (i.e., $k = 3$ in our case) to "nominative-accusative," each of which is supposed to be calculated by correspondence analysis. Here, we assume that A language is plotted as 0.95, 0.9, 0.85 in each dimension respectively, B language as 0.9, 0.85, 0.55 in each dimension respectively, and C language as 0.7, 0.75, 0.85 in each dimension respectively.



Ainu = Chukchi + Khalkha + Navajo + Ainu error
Chukchi = Ainu + Khalkha + Navajo + Chukchi error
Khalkha = Ainu + Chukchi + Navajo + Khalkha error
Navajo = Ainu + Chukchi + Khalkha + Navajo error

Figure 2. A basic image of the relationship between one language and the other languages in linguistic typology. Blue directional arrows correspond to the correlation between languages in linguistic typological characteristics.

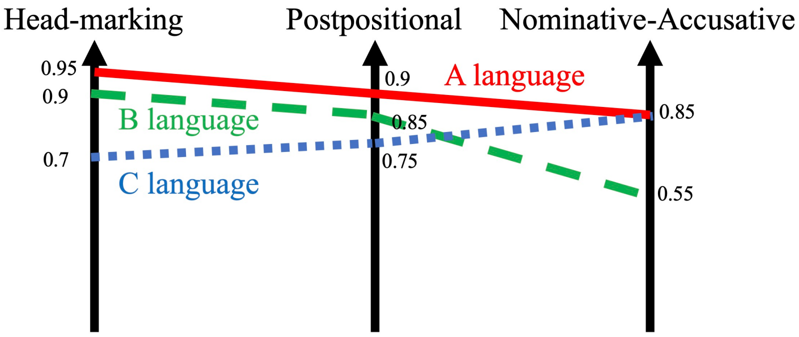


Figure 3. Example of quantification obtained by correspondence analysis. The first dimension, where k as 1 in $\sqrt{\omega_k}x_{ik}$ tentatively corresponds to the degree of head-marking; the second dimension, where k as 2 in $\sqrt{\omega_k}x_{ik}$ corresponds to the degree of postpositional; the third dimension, where k as 3 in $\sqrt{\omega_k}x_{ik}$ corresponds to the degree of nominative-accusative, respectively. $i = 1$ (A language), $i = 2$ (B language), and $i = 3$ (C language) for example.

If linguistic typologists consider similarity or dissimilarity between languages from the viewpoints of "distance," then distance can be |0.95 - 0.9| + |0.9 - 0.85| + |0.85 - 0.55| = 0.4 between A language and B language, and distance can be |0.95 - 0.7| + |0.9 - 0.75| + |0.85 - 0.85| = 0.4 between A language and C language. Thus, A language is similar to both B language and C language in linguistic typology from this "distance view."

However, this "distance view" is counter-intuitive in linguistic typology because most linguistic typologists will consider B language more similar to A language and C language is less similar to A language.

In this paper, we will call this viewpoint a "profile view" in linguistic typology and propose that the Pearson product-moment correlation coefficient (abbreviated correlation hereafter) realize a "profile view" in numerical vectors on each language calculated by correspondence analysis. The estimated correlation (i.e., r_{ij}) between languages is:

$$r_{ij}=\frac{\sum_{m=1}^{k}\left\{\left(\sqrt{\omega_m}x_{im}-\frac{1}{k}\sum_{l=1}^{k}\sqrt{\omega_l}x_{il}\right)\cdot\left(\sqrt{\omega_m}x_{jm}-\frac{1}{k}\sum_{l=1}^{k}\sqrt{\omega_l}x_{jl}\right)\right\}}{\sqrt{\sum_{m=1}^{k}\left(\sqrt{\omega_m}x_{im}-\frac{1}{k}\sum_{l=1}^{k}\sqrt{\omega_l}x_{il}\right)^2\cdot\sum_{m=1}^{k}\left(\sqrt{\omega_m}x_{jm}-\frac{1}{k}\sum_{l=1}^{k}\sqrt{\omega_l}x_{jl}\right)^2}}$$

Euclidean distance and Manhattan distance between languages are:

$$d\_E_{ij}=\sqrt{\sum_{m=1}^{k}\omega_m\cdot\left(x_{im}-x_{jm}\right)^2},\ d\_M_{ij}=\sum_{m=1}^{k}\sqrt{\omega_m}\cdot\left|x_{im}-x_{jm}\right|$$

For example, the estimated correlation is calculated as 0.92 between A language and B language, as -0.98 between A language and C language in Figure 3, clarifying that the B language is more similar to A language than C language is as consistent with some intuition in linguistic typology. Thus, correlation can represent a "profile view" in linguistic typology.

Since correlation can vary from -1 to 1, in other words, correlation is not an interval scale but an ordinal scale, we cannot apply statistical classification methods based on "distance." Instead, we will apply Partitioning Around Medoids or K-medoids (Kaufman and Rousseeuw 1990; abbreviated as PAM hereafter) because PAM is applicable for grouping data based on ordinal similarity (i.e., correlation in our case). Since it is difficult from the viewpoints of computation that our results in PAM are represented as dendrogram and other clustering methods are more useful for evaluating our viewpoints of correspondence analysis and correlation similarity to areality and genealogy in linguistic typology. Thus, we will also apply the Ward method (Ward 1963) and complete method (Sørensen 1948) to correlation similarity between languages in our data.

5.3. Linguistic Requirement (3): Graphical Modeling Approach

The previous subsection showed that correlation is admissible for "profile view" in linguistic typology. In this subsection, we will aim to measure both areality and genealogy in linguistic typology simultaneously, using the correlation coefficient.

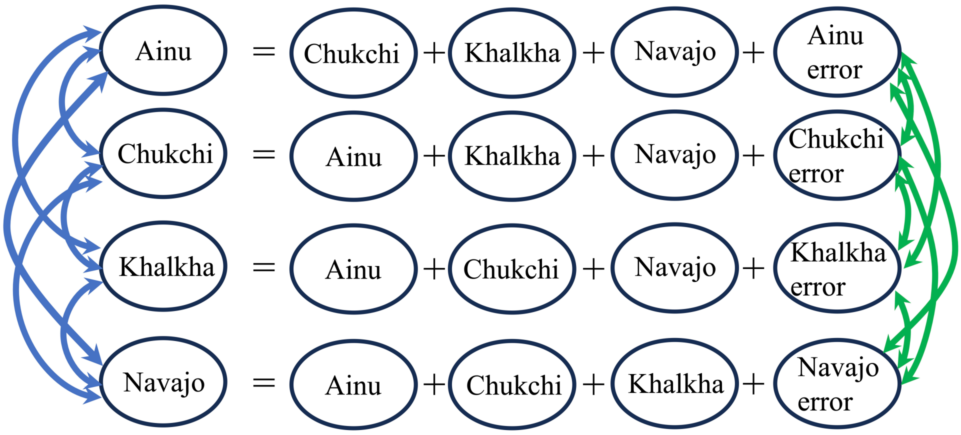


Figure 4. The basic image of the relationship between one language and the other languages in linguistic typology. Blue directional arrows correspond to a correlation between languages in linguistic typological characteristics, and green directional arrows correspond to a correlation between languages that cannot be explained by linguistic typological characteristics (e.g., geographical factors).

As introduced in Section 4, "linguistic typological characteristics in language A" (i.e., "language A") is comprised of two notations: "linguistic typological characteristics in language A that can be explained by linguistic typological character in the other languages" and "linguistic typological characteristics in language A that cannot be explained by linguistic typological characteristics in the other languages" (i.e., "language A error"), the latter of which can be considered as linguistic typological characteristics explained by geographical factors.

Since our main concerns in this paper are to capture both areality and genealogy in linguistic typology simultaneously, new statistical methods are required for analyzing the correlation not only between "language A" and "language B" but also between "language A error" and "language B error," latter of which corresponds to green bidirectional arrows in right side on Figure 4.

Based on "language as variable", we will build multiple linear regression model[5]:

$$\sqrt{\omega_m}x_{1m} = \beta_{01} + \beta_{21}\sqrt{\omega_m}x_{2m} + \beta_{31}\sqrt{\omega_m}x_{3m} + \beta_{41}\sqrt{\omega_m}x_{4m} + \varepsilon_{1m}$$

$$\sqrt{\omega_m}x_{2m} = \beta_{02} + \beta_{12}\sqrt{\omega_m}x_{1m} + \beta_{32}\sqrt{\omega_m}x_{3m} + \beta_{42}\sqrt{\omega_m}x_{4m} + \varepsilon_{2m}$$

$$\sqrt{\omega_m}x_{3m} = \beta_{03} + \beta_{13}\sqrt{\omega_m}x_{1m} + \beta_{23}\sqrt{\omega_m}x_{2m} + \beta_{43}\sqrt{\omega_m}x_{4m} + \varepsilon_{3m}$$

$$\sqrt{\omega_m}x_{4m} = \beta_{04} + \beta_{14}\sqrt{\omega_m}x_{1m} + \beta_{24}\sqrt{\omega_m}x_{2m} + \beta_{34}\sqrt{\omega_m}x_{3m} + \varepsilon_{4m}$$

[5] $m = 1, 2, \ldots, n-1$, and $i = 1$ (Ainu), $i = 2$ (Chukchi), $i = 3$ (Khalkha), $i = 4$ (Navajo) in our example.

Our focuses are not only r_{ij} corresponding to blue bidirectional arrows in Figure 4 but also the estimated correlation between residuals (i.e., between the estimated ε_{im} and the estimated ε_{jm}, $i, j = 1, 2, \ldots 4, i \neq j$ in Figure 4) corresponding to green bidirectional arrows in Figure 4.

Fortunately, previous studies (Tsubaki and Tsubaki 1997; Tsubaki 2011a, 2011b) have already proposed how to analyze blue and green bidirectional arrows in correlation analysis simultaneously based on some motivation in Anderson (1984). Thus, Tsubaki (2011a, 2011b) has demonstrated that partial residual correlation (i.e., green bidirectional arrows on the right side in Figure 4) coincided with the inverse correlation matrix in our data. Moreover, the correlation matrix and inverse correlation matrix have the same eigenvectors, but the reciprocal of eigenvalues in the correlation matrix corresponds to the eigenvalues in the inverse correlation matrix, the latter of which will play a significant role in graphical modeling (Lauritzen 1996)[6].

We determined to adopt the ratio of change rate of eigenvalues[7] obtained by correspondence analysis as an empirical criterion for distinguishing "meaningful" dimensions from "noise" dimensions because statisticians often have utilized the dimension illustrating a high ratio of change rate of eigenvalues as "elbow," and selected up to the dimensions before "elbow" as meaningful dimensions (Greenacre 2017).

From the viewpoint of statistics, Arai et al. (2001) have already applied Tsubaki's (2011a, 2011b) method, but our method is novel in applying Tsubaki's method to the correlation matrix between subjects (i.e., languages in our case) instead of correlation matrix between variables (i.e., features in our case), integrating the three linguistic requirements consistent to some assumptions underlying statistical methods.

Thus, our basic framework in statistical analyses is summarized as follows: (1)

[6] Furthermore, in correlation analysis, Tsubaki and Tsubaki (1997) have already proposed how to distinguish substantive meaningful dimension from "noise" dimension using criteria of deviance as to selecting "noise" dimensions whose eigenvalue will minimize the coefficient of variation in eigenvalues or variance of log of eigenvalue. Since traditional dimension selection in principal component analysis (Jolliffe 2002) will emphasize the dimension with a larger eigenvalue and graphical modeling will focus on the dimension with a smaller eigenvalue, Tsubaki and Tsubaki (1997) and Tsubaki (2011a, 2011b) have integrated both of principal component analysis and graphical modeling regarding dimension selection. Their alternatives are to consider the dimension with intermediate eigenvalue as the "noise" dimension from the viewpoints of deviance. Thus, Arai et al. (2001) have applied Tsubaki's statistical methods to Alzheimer's disease data and brought new insights into cognition functions with different Alzheimer types. However, we cannot directly apply the previous method (Tsubaki and Tsubaki 1997; Tsubaki 2011a, 2011b) to our data because they have assumed the data to be distributed by multivariate normal distribution that is valid in the case of both principal component analysis and graphical modeling. Since correspondence analysis in this paper will apply to binary categorical data and transform the data into continuous data using chi statistic as introduced in Section 5.1, we cannot assume the data to be distributed by multivariate normal distribution. Thus, we will leave further statistical discussions to another article in the future.

[7] The ratio of the change rate of eigenvalues in the ith dimension is defined on Figure 9 in Section 7.2 as $(\omega_{i+2}/\omega_{i+1})/(\omega_{i+1}/\omega_i)$

transform the original data in WALS into our proposed format; (2) apply correspondence analysis to (1); (3) calculate correlation matrix based on numerical vector by (2); (4) plot log of eigenvalues on correlation matrix; (5) from viewpoints of ratio of change rate of eigenvalues, select dimension including largest eigenvalues and smallest eigenvalues, and exclude "noise" dimensions containing intermediate eigenvalues; (6) recalculate correlation matrix based on selected dimension by (5); (7) apply PAM or Ward clustering techniques to correlation matrix by (6).

6. Evaluations of Clustering

In this section, we will deal with the evaluation of clustering. One of the purposes of this paper is to consider the areality in the typological dataset revealed in Whitman and Ono (2015) and Ono and Whitman (2016) in more numerical ways based on hierarchical cluster analysis. In hierarchical cluster analysis, how many clusters to choose is always problematic.

In this paper, some statistical criterion to evaluate the number of clusters is needed. We first automatically determine the number of clusters in each analysis using Krzanowski-Lai Index (Krzanowski and Lai 1988) and evaluate the classification based on the areality we assume by applying the Rand Index (Rand 1971). Each statistical analysis is implemented in R language (R Core Team 2023).

6.1 Determining the Optimal Number of Clusters

There are many criteria to determine the number of clusters in both hierarchical and non-hierarchical clustering. In this paper, the Krzanowski-Lai Index is selected as a tentative approach[8]. In R languages, the "NbClust" package (Charrad, Ghazzali, Boiteau, and Niknafs 2014) implemented this index. We use the "NbClust" package to determine the number of clusters automatically, with the minimum number of clusters as 20 and the maximum as 50.

6.2 Evaluation of Clustering Under the Optimal Number of Clusters

After determining the number of clusters, we calculate the goodness of the classification with respect to the areality in linguistic typology using the Rand Index, which is implemented by the "phyclust" package (Chen and Dorman 2010) in R language. The explanation about the Rand Index is as follows: Let π be the "true" classification and ρ be one result of clustering. We define a_{11} as the numbers of a combination of two languages, which is the same cluster in both π and ρ; a_{12} as those which are the same cluster in π but not in ρ; a_{13} as those which are not the same cluster in π but in ρ;

[8] Krzanowski-Lai Index uses the trace information of the within-group dispersion matrix to optimize the number of clusters. We do not mention here in detail about Krzanowski-Lai Index and only to refer Tibshirani, Walther, and Hastie (2001).

a_{14} as those, which is neither the same cluster in π nor in ρ. Thus, the Rand Index is defined as follows;

$$Rand\ Index = \frac{a_{11} + a_{14}}{a_{11} + a_{12} + a_{13} + a_{14}} = \frac{a_{11} + a_{14}}{\binom{N}{2}}$$

, where *N* is the number of languages (i.e. *N* is 201 in our case of Data1-3).

The larger the Rand Index is, the more appropriate the classification in terms of areality is. The “true” classification of 201 languages is given linguistically in advance, and then we automatically compare the classification of the output of “NbClust” and calculate the Rand Index using the “phyclust” package in R language.

There are 5 languages in the Ancient Near East, 32 in Africa, 15 in Australia, 7 in East North America, 18 in Europe, 12 in Mesoamerica, 10 in North Asia, 19 in New Guinea, 29 in South America, 35 in South and South East Asia, and 19 in West North America. Interested readers can refer to the “true” classification of 201 languages in Ono and Whitman (2016).

7. Results

In this section, we will first verify whether previous studies (Whitman and Ono 2015; Ono and Whitman 2016) have revealed areality and genealogy in linguistic typology just as a coincidence or their analyses based on correspondence analysis and correlation similarity, which correspond to “correction on frequencies” and “profile view” respectively, can be supported from the results of Rand Index. Subsequently, we will also examine whether “language as variable” can improve classification results in determining “noise” dimensions and selecting “meaningful” dimensions, applying Tsubaki’s (2011a, 2011b) viewpoints.

7.1 Results on Data 1-3: Validating “correction on frequencies” and “profile view”

This subsection shows our results on statistical analyses applied to Data 1-3. First, the Rand index applied to Data1-3 is summarized in Figure 5. Figure 5 illustrates the Rand index as a boxplot in terms of (dis)similarity measure and clearly shows that areality and genealogy are more detected in clustering results using correlation similarity, compared to Euclidean or Manhattan distance, which resulted in supporting our viewpoint of “profile view” in linguistic typology, instead of “distance view” in Section 5.2.

Furthermore, Figure 6 illustrates the Rand index as a boxplot in terms of clustering techniques and clearly shows that areality and genealogy are more detected in clustering results using the Ward method, compared to the complete method, which resulted in supporting our findings in Whitman and Ono (2015) and Ono and Whitman (2016) in more numerical ways. From the viewpoints of the graphical modeling approach in Section 5.3, the results on the Rand index are summarized as to three points: quantification methods, dimension selection, and similarity measure in Figure 7.

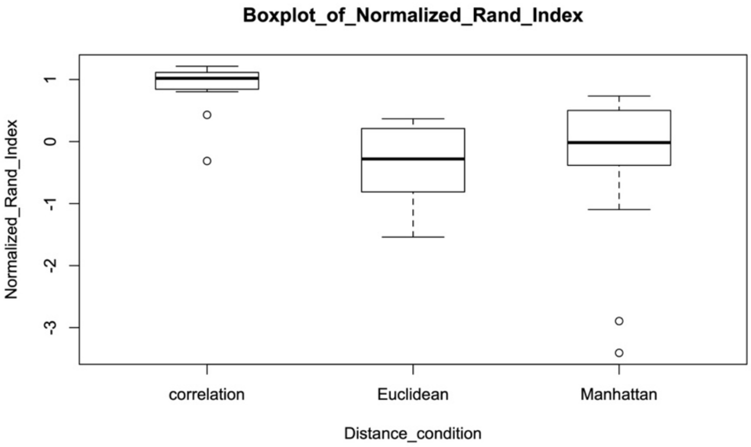


Figure 5. Boxplots of normalized Rand Index in each distance condition. Correlation (i.e., “profile view in our case”) outperformed in normalized Rand Index, compared to Euclidean and Manhattan distance (“distance view” in our case).

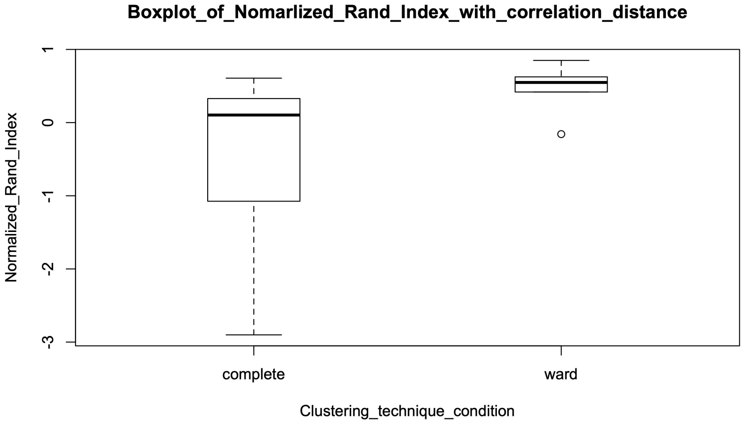


Figure 6. Boxplots of normalized Rand Index in each clustering technique condition. Ward method (ward) used in previous studies (Whitman and Ono 2015; Ono and Whitman 2016) outperformed in normalized Rand Index compared to the complete method (complete).

Clustering results using correspondence analysis with all dimensions and correlation similarity (i.e., MCA_full_correlation) outperformed those including only dimensions with larger eigenvalues and Euclidean and Manhattan distance (i.e., MCA_tandem_Euclidean_Manhattan), which resulted in partially supporting our “language as variable,” including some dimensions with smaller eigenvalues in statistical analyses[9].

[9] In conditions of Figure 7 with MCA, tandem, and correlation, the Rand Index cannot be calculated by chaining (Everitt, Landau, Leese, and Stahl 2011), or the dendrogram was collapsed. Thus, we cannot directly compare MCA_full_correlation to MCA_tandem_correaltion in Rand Index.

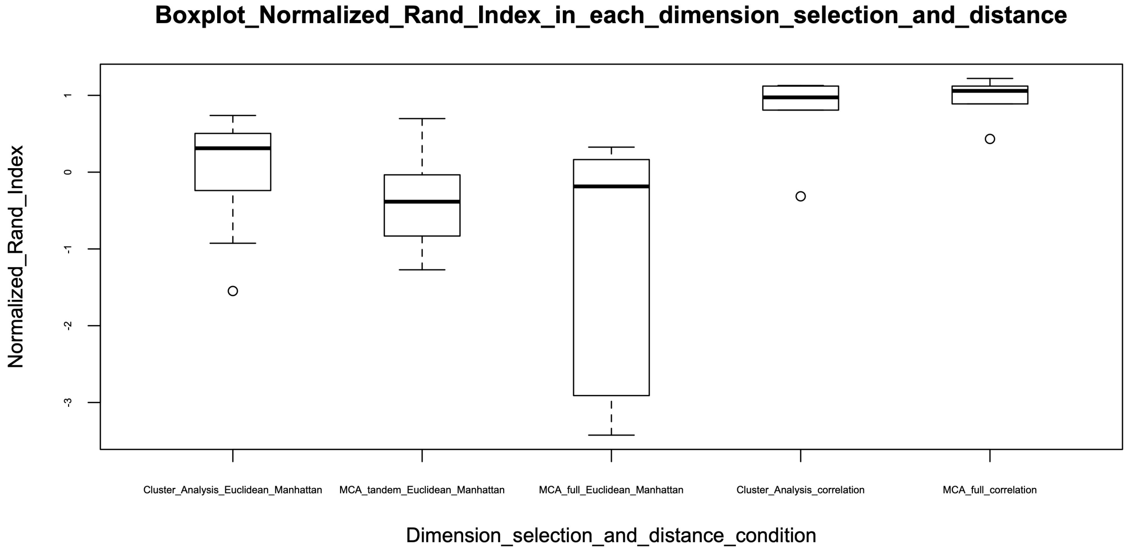


Figure 7. Boxplots of normalized Rand Index in each clustering condition about quantification, dimension selection, and similarity measure. Quantifications have two conditions: Cluster analysis (utilize raw binary categorical data in WALS); and MCA (apply correspondence analysis to data). Dimension selections in MCA have three conditions: tandem (using only dimensions with larger eigenvalues like principal component analysis); and full (using all dimensions including smaller eigenvalues like graphical modeling). The similarity measure is a correlation, Euclidean or Manhattan.

Moreover, clustering results using correspondence analysis with all dimensions and correlation similarity (i.e., MCA_full_correlation) outperformed those using just correlation similarity (i.e., Cluster_Analysis_correlation) slightly, which resulted in supporting the viewpoint of "correction on frequencies" in Section 5.1 because correlation similarity is statistically valid in the Euclidean coordinates obtained by correspondence analysis but invalid in the configuration of binary data[10].

Thus, it is numerically verified in this subsection that the previous studies (Whitman and Ono 2015; Ono and Whitman 2016) applying MCA_full_correlation to Data 3 have revealed areality and genealogy in WALS data not just as coincidence but based on the background of both linguistics and statistics. Thus, the next subsection will examine whether our viewpoint of "language as variable" can improve the finding on areality and genealogy in linguistic typology.

7.2 Results on Data 4: Validating "language as variable"

Our viewpoints of "correction on frequencies" and "profile view" have been supported

[10] In other words of statistics, correlation similarity does not work theoretically when applied to binary data (i.e., raw binary categorical WALS data in our case) because correlation similarity assumes Euclidean space. However, the Rand Index did not deteriorate in Cluster_Analysis_correlation. Thus, further investigation of this finding could contribute to the theoretical part of statistics.

by the results in the previous subsection. In this subsection, we will validate whether our viewpoint of "language as variable" works in detecting areality and genealogy in linguistic typology on Data 4.

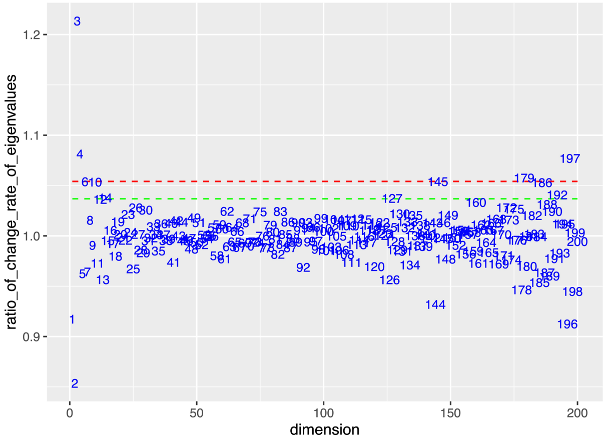


Figure 8. Plot of ratio of the change rate of eigenvalues (ra_i) obtained by correspondence analysis to Data 4. The ratio of the change rate of eigenvalues is defined as $ra_i = (\omega_{i+2}/\omega_{i+1})/(\omega_{i+1}/\omega_i)$.

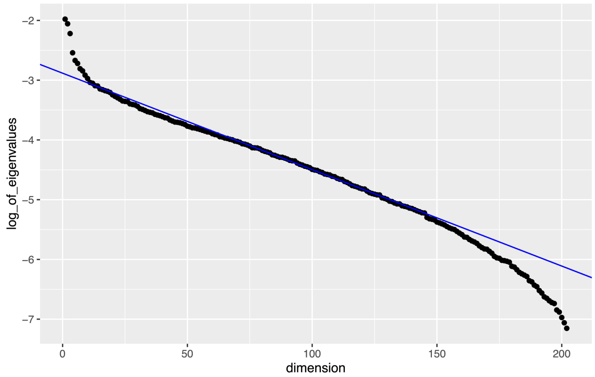


Figure 9. The plot of the log of eigenvalues obtained by correspondence analysis to Data 4. The blue line is plotted from the 13th dimension to the 145st dimension.

We determined the coordinates, which were obtained by correspondence analysis from the 13th to the 145th dimension as the "noise" dimension, and selected the coordinates from the 1st to the 12th dimension and the 146th to 202nd dimension in the following analyses[11].

[11] Figure 8 illustrates the ratio of the change rate of eigenvalues obtained by correspondence analysis to Data 4. Here, "elbow" corresponds to the 13th dimension because $ra_{12} = (\omega_{14}/\omega_{13})/(\omega_{13}/\omega_{12})$ showed a positive value over 1 in Figure 8. Thus, we have tentatively determined to include the 1st to 12th dimension before the elbow and from the 146th dimension to the 202nd dimension as "meaningful" dimensions because the ratio of the change rate of eigenvalues does not vary differently up to 145th dimension and $ra_{145} = (\omega_{147}/\omega_{146})/(\omega_{146}/\omega_{145})$ showed positive value over 1. Furthermore,

We have applied PAM to the coordinates obtained by correspondence analysis that corresponds to C1: from 1st to 12th dimension and from 146th to 202nd dimensions; C2: all dimensions (from 1st to 202nd dimension); C3: from 1st to 12th dimension; C4: From 1st to 145th dimension; C5: From 146th to 202nd dimension; C6: From 13th to 202nd dimension in Table 16 on supplementary materials, respectively. Since pamk function in the "fpc" package (Hennig 2023) calculated the optimal number of clustering as 7 in C1, we have chosen the number of clustering as 7 from C1 to C6 in supplementary materials.

As demonstrated in supplementary materials, the classification results in C5 and C6 cannot be interpreted from areality and genealogy in language, which resulted in the significance of the coordinates from the 1st to 12th dimension. Furthermore, comparing C4 to C2, the classification results improved areally and genetically in C2, which supported our viewpoints of "graphical modeling," emphasizing dimensions with smaller eigenvalues (i.e., from 146th to 202nd dimension in our case). The same analysis can apply to the classification results in C1, showing the areality in languages in Africa, compared to those in C3, with the information from the 146th to 202nd dimension.

Finally, we can identify in the classification result in C1 the "noise" dimensions as 13th to 145th dimension with intermediate eigenvalues, compared to those in C2. The classification result in C2 resulted in not only containing languages in Africa into Circum-Pacific structure (Nichols 1994; Bickel and Nichols 2006) corresponding to 1-3 in cluster id of C1 but also clustering both Indo-European languages and African languages into one group in C2 (i.e., cluster id 6).

Thus, from the viewpoint of statistics, these classifications resulted in supporting that our view of "language as variable" in Section 5.3 improved areality and genealogy in linguistic typology by graphical modeling approach integrating dimensions with larger and smaller eigenvalues in correlation analysis through a new perspective of previous studies (Tsubaki and Tsubaki 1997; Tsubaki 2011a, 2011b). From the viewpoint of linguistics, we illustrated the classification results in C1 and C3 as the map of Figure 10 and Figure 11, respectively, utilizing MATLAB (The MathWorks Inc. 2022) to cluster id in C1 (i.e., C1_id) and C3 (i.e., C3_id). In Figure 10, the hierarchical structure can be denoted as {{{1, {2, 3}}, {{4, 5}, 6}}, 7}; {2, 3} shows clusters 2 and 3 form one cluster, and the same analysis applies to other notations. Thus, Circum-Pacific structure (Nichols 1994, Bickel and Nichols 2006) corresponding to circles in Figure 10 is detected as {1, {2, 3}}, and cluster 4 corresponds to languages in Africa, cluster 5 to those in Eurasia, cluster 6 to those in Europe or Indo-European family, and cluster 7 to those in Austronesian and Austro-Asiatic with some exceptional languages.

Figure 9 plots a log of eigenvalues obtained by correspondence analysis to Data 4. As shown in the blue line in Figure 9, the log of eigenvalues decreases by a nearly equal degree from the 13th dimension to the 145th dimension that the previous studies (Tsubaki and Tsubaki 1997; Tsubaki 2011a, 2011b) have recommended selecting "noise" dimensions as correlation analysis.

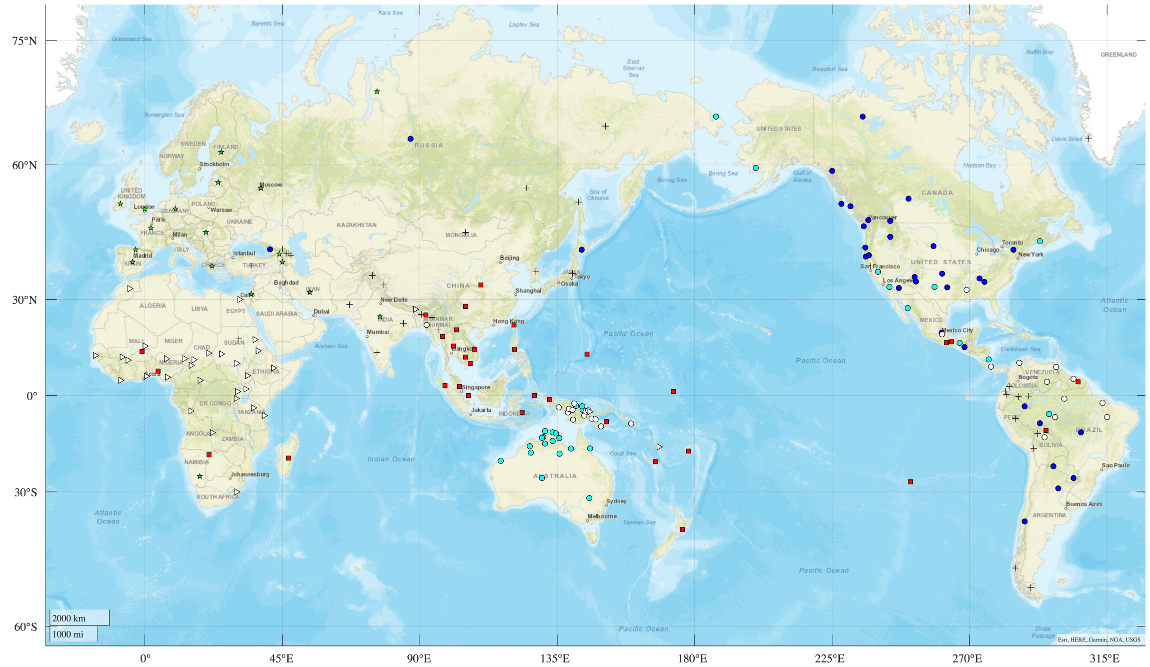


Figure 10. The classification results in C1 (i.e., from 1st to 12th dimension and from the 146th to 202nd dimensions). Blue circle: cluster 1; cyan circle: cluster 2; white circle: cluster 3; triangle: cluster 4; plus: cluster 5; star: cluster 6; rectangle: cluster 7 in C1 in Table 16 on supplementary materials.

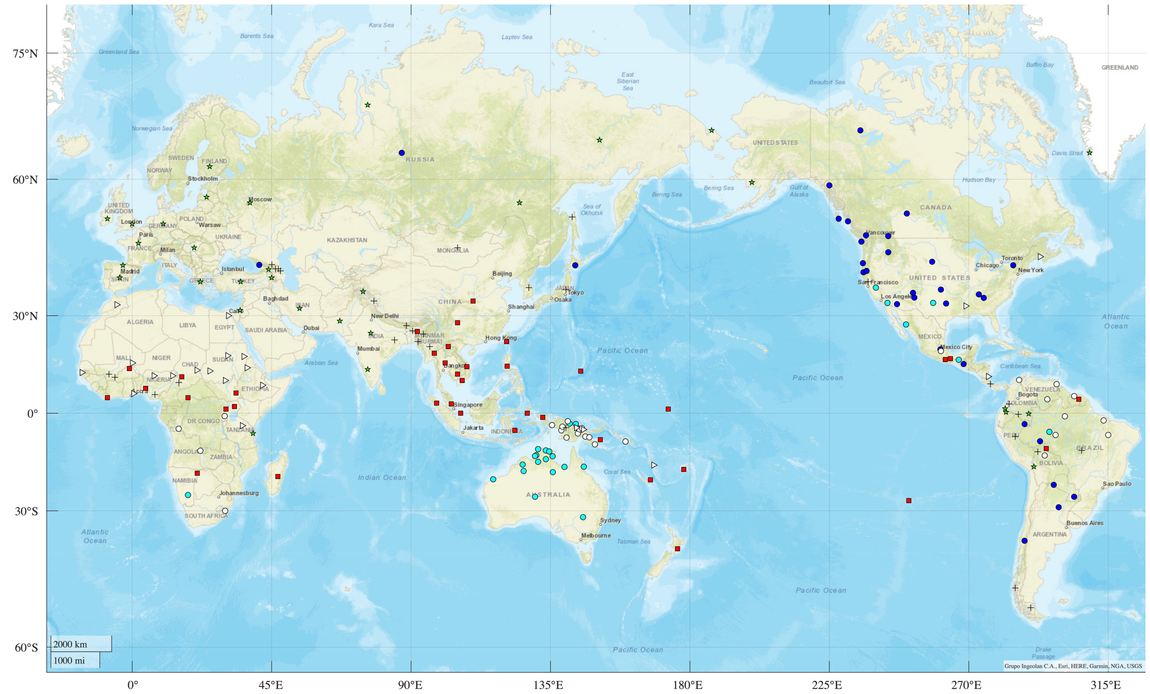


Figure 11. The classification results in C3 (i.e., from the 1st to the 12st dimension). Blue circle: cluster 1; cyan circle: cluster 2; white circle: cluster 3; triangle: cluster 4; plus: cluster 5; star: cluster 6; rectangle: cluster 7 in C3 in Table 16 on supplementary materials.

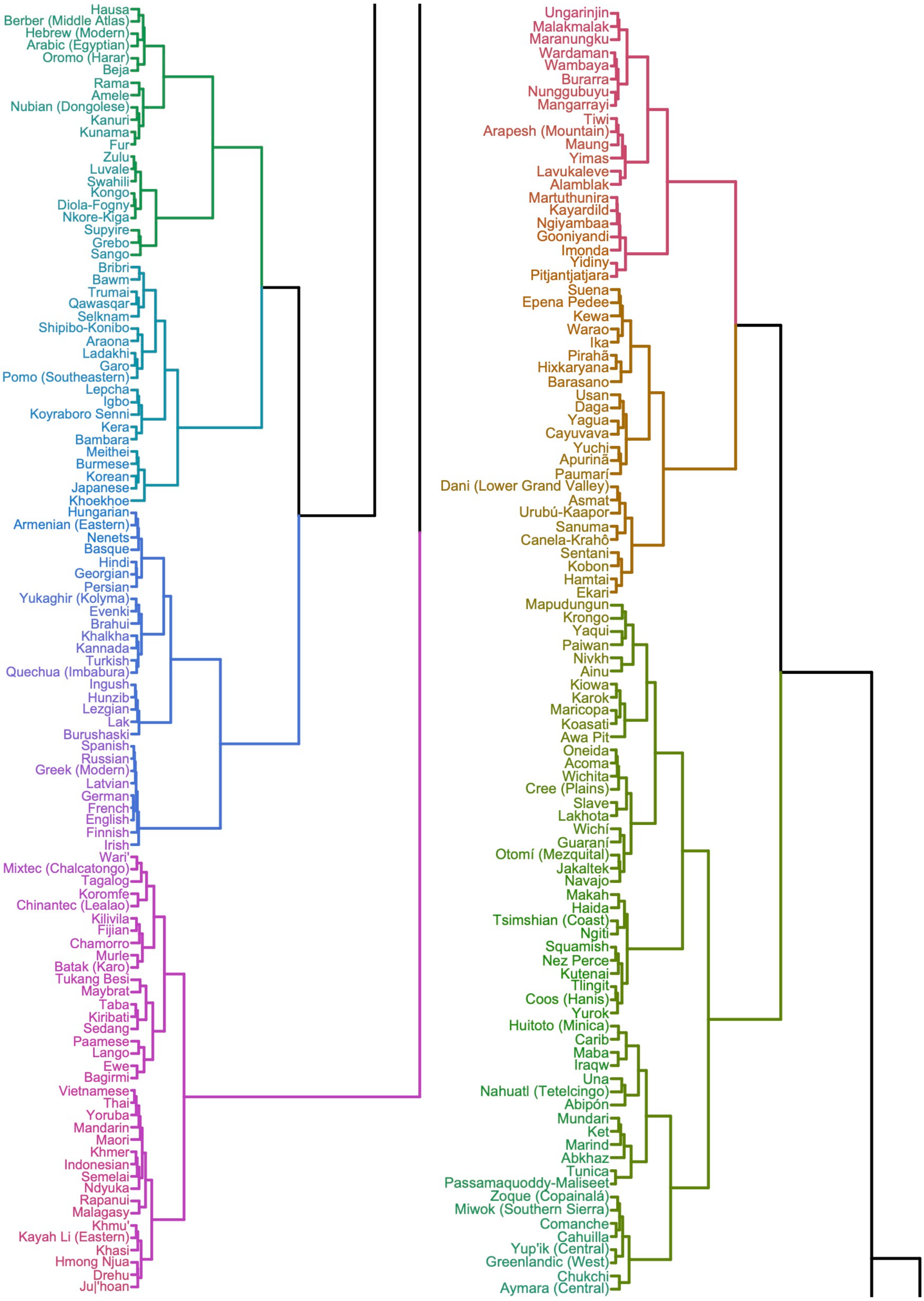


Figure 12. The hierarchical clustering result in C1 utilizing Ward method (Ward 1963). Right: The upper part of the clustering result; Left: The lower part of clustering result.

In Figure 11, the Circum-Pacific structure is more violated by languages including Africa (triangle in Figure 11), Eurasia (plus in Figure 11), Europe, or Indo-European (star in Figure 11) languages. Furthermore, areality in Africa does not hold by mixing the cluster in Austronesian and Austro-Asiatic, and vice versa.

Moreover, Figure 12 illustrated dendrogram obtained from Data 4 with Ward method and correlation coefficients in Figure 10. We can observe not only specific clusterings on part of language family (e.g., Afro-Asiatic language family, Niger-Congo language family, North-Caucasian language family, Indo-European language family, Australian languages), but also pairs of languages (e.g., Ainu and Nivkh, Japanese and Korean, Yupi'k [Central] and Greenlandic [West]) in Figure 12, which were not clearly captured in Figure 10. Thus, we will leave another research to further discussions on genealogical relationships illustrated in this paper.

Thus, from the viewpoint of linguistics, our view of "language as variable" improved previous studies (Ono and Whitman 2016) on linguistic typology, not only about areality and genealogy in Africa, Eurasia, in Europe or Indo-European families, or Austronesian and Austro-Asiatic but also Circum-Pacific structure in languages.

8. Discussions and Conclusions

This section discusses the significance of this paper from the perspectives of both linguistics and statistics. From the viewpoint of linguistics, the main results can be summarized in four points. First, our classification results illustrated areality and genealogy in linguistic typology based on the data without word-order features that previous studies have found as highly correlated with each other, which suggests the word-order features have been masking areal and genetic information in linguistic typology and led previous research to some labyrinth in linguistic typology. Thus, further linguistic analyses will provide significant relationships between word-order parameters and areal and genetic information in languages.

Second, Data 3 and Data 4 have included word-order features related to Adjectives and Negation in WALS and resulted in improving areality and genealogy as a result. Thus, Adjectives and Negation should be worth studying not only in linguistic universals (Greenberg 1963) but also in a more general context (Chomsky 2014), considering differences to other word-order features.

Third, this paper illustrates areality and genealogy in linguistic typology, including Circum-Pacific structure in previous studies (Nichols 1994; Bickel and Nichols 2006). Since this paper tentatively assumed that our data are sufficiently representative of languages and features in linguistic typology, there will be some possibility that the data revised for such representativeness will improve our classification results.

Finally, the Circum-Pacific structure detected in this paper includes some Pacific Rim. However, our classification result did not distinguish the coastal part of languages from the other part in North and South America, the former of which strictly consist of the Pacific Rim (Bickel and Nichols 2006: 6). Thus, further detailed features will potentially reveal not only Pacific Rim structure but also North Pacific Rim (Miyaoka 1992).

Thereby, our result suggests that improving the quality of the database in linguistic typology is still promising in addressing areality and genealogy in languages that cannot be traced back by another realm, including lexicostatistics[12].

From the viewpoints of statistics, the main results can be summarized in three points. First, our classification results showed the answer to unresolved problems in Ono (2020): whether we can approach areality and genealogy of languages globally by linguistic typology without imputing missing values in linguistic typology. Our proposed statistical analyses do not require statistical imputation and enable the researcher to address areality and genealogy in linguistic typology.

Second, as shown in our results, our three viewpoints have played a significant role in the statistical analysis as a result: "correction on frequencies," "profile view," and "language as variable," the last of which will develop novel methodologies not limited in linguistics but in the huminites as whole. Since documents are usually related to each other, or documents influence other documents in the humanities, our viewpoint of "language as variable" can extend to "document as variable;" documents can be considered as a function of other documents. Thus, graphical modeling or correlation analysis utilizing eigenvectors with both larger and smaller eigenvalues will be promising as exploratory analysis in the humanities.

Third, this paper revealed novel statistical issues in dimension selections on correspondence analysis. As shown in Figure 10, the classification result is improved in areal information by including dimensions with smaller eigenvalues that correspond to some sort of correlation between partial residuals or inverse correlation related to the actual similarity between languages not explained by linguistic typology (e.g., geographical factors). However, as explained in Section 5.3, previous theoretical studies (Tsubaki and Tsubaki 1997; Tsubaki 2011a, 2011b) have focused on the case where researchers will apply principal component analysis and need to perform dimension selection mainly on variables, under which we can assume the data to be distributed by normal distribution. As correspondence analysis cannot assume the data following normal distribution and our proposed method will consider the correlation between subjects (i.e.,

[12] Moreover, one of reviewers suggested further developments of this paper. Since areal linguistics has mainly been focusing on some features or values that are not found in other areas or not necessarily included in WALS, there are some limitations in applying the results of WALS to areal linguistics. However, this paper illustrated with some objective methods that the common linguistic typological features or values in WALS can deal with areality in linguistics (e.g., Circum-Pacific clustering), which will potentially bright new methodologies to present areal linguistics that has been capturing areal features or values not in statistical sense.

languages in our case), further statistical studies are needed on how to construct clear criteria (i.e., deviance in Tsubaki and Tsubaki [1997]) for determining "noise" dimensions from a theoretical point of view.

In conclusion, our statistical analyses combining correspondence analysis, correlation analysis, graphical modeling, and clustering techniques are based on scrutinizing and satisfying three linguistic requirements, which resulted in the improvement of areality and genealogy in linguistic typology. Thus, these facts suggest that interdisciplinary research should incorporate different perspectives in each realm into one idea consistent with how researchers understand the phenomenon. This will result in new insights and new methodologies in each field.

Conversely, any interdisciplinary research without the above conditions will lead to endless labyrinth in statistical analyses. We end this paper in the hope that, in interdisciplinary research, it will be rightly recognized the correspondence between mathematical assumptions in statistical methods and assumptions in humanities data.

Acknowledgments

I am very grateful to Professor John Whitman (Cornell University/ Emeritus Professor of the National Institute for Japanese Language and Linguistics) for our collaborative research. Parts of this paper were presented at the 5th annual conference of The Japan Association of Northern Language Studies, on November 26, 2022. I am sincerely grateful to the conference organizers and participants, especially Professor Hirofumi Hori (Shizuoka University), Professor Fuyuki Ebata (Niigata University), and Professor Norikazu Kogura (Tokyo University of Foreign Studies), for their invaluable questions and comments. Finally, I would like to thank the editors and two highly conscientious reviewers for their constructive and invaluable comments; all errors are of course my own.

References

Anderson, Theodore Wilbur (1984) *An introduction to multivariate statistical analysis,* Second Edition. New York: Wiley.

Arai, Heii, Hiroe Tsubaki, Yoshio Mitsuyama, Naoki Fujimoto, Yasuo Urata, and Akira Homma (2001) Early onset Alzheimer type dementia more rapidly deteriorates than late-onset type: A follow-up study on MMSE scores in Japanese patients, *Psychogeriatrics*, 1: 303–308.

Benzécri, Jean-Paul et coll. (1973) *L'analyse des données. Volume II: L'analyse des correspondances*. Paris: Bordas.

Bickel, Balthasar and Johanna Nichols (2006) Oceania, the Pacific Rim, and the theory of linguistic areas. *Annual Meeting of the Berkeley Linguistics Society*, 32(2): 3–15.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and

John B. Lowe (2017) The AUTOTYP typological databases. Version 0.1.0

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs (2014) NbClust: an R Package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6): 1–36. URL http://www.jstatsoft.org/v61/i06/

Chen, Wei-Chen and Karin Dorman (2010) phyclust: Phylogenetic clustering (Phyloclustering). R package, URL http://cran.r-project.org/package=phyclust

Chomsky, Noam (2014) *The minimalist program*. 20th annual edition, MIT press.

Corbett, Greville G. (2013) Number of genders. 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/30, Accessed on 2023-12-06.)

Daumé III, Hal (2009) Non-parametric Bayesian areal linguistics. In Johnston, Michael and Popowich Fred (eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 593–601. Boulder: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N091067.pdf

Daumé III, Hal and Lyle Campbell (2007) A Bayesian model for discovering typological implications. In Zaenen, Annie and van den Bosch Antal (eds.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 65–72. Prague: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P07-1009.pdf

Drummond, Alexei, Marc Suchard, Dong Xie, and Andrew Rambaut (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8): 1969–1973.

Dryer, Matthew S. (1992) The Greenbergian word-order correlations. *Language*, 68(1): 81–138.

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.)

Everitt, Brian. S., Sabine Landau, Morven Leese, and Daniel Stahl (2011) *Cluster analysis.* New York: Wiley.

Goedemans, Rob and Harry van der Hulst (2013) Fixed stress locations. 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 2023-12-06.)

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

Greenberg, Joseph (1963) Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, Joseph (ed.), *Universals of language*, 73–113. London: MIT Press.

Hennig, Christian (2023) fpc: Flexible procedures for clustering. R package version 2.2-10, URL https://CRAN.R-project.org/package=fpc

Hymes, Dell H. (1960) Lexicostatistics so far. *Current Anthropology*, 1(1): 3–44.

Jambu, Michel (1989) *Exploration informatique et statistique des données.* Paris: Dunod.

Jolliffe, Ian T. (2002) *Principal component analysis.* New York: Springer New York.

Kaufman, Leonard and Peter J. Rousseeuw (1990) Partitioning around medoids (Program PAM). In *Finding groups in data: An introduction to cluster analysis*, 68–125, New York: John Wiley & Sons, Inc.

König, Ekkehard, Peter Siemund, and Stephan Töpper (2013) Intensifiers and reflexive pronouns. 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/47, Accessed on 2023-12-06.)

Krzanowski, Wojtek J., and Y. T. Lai (1988) A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1): 23–34.

Lauritzen, Steffen L. (1996) *Graphical models*. Oxford: Clarendon Press.

Maddieson, Ian (2013) Syllable structure. 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/12, Accessed on 2023-12-06.)

The MathWorks Inc. (2022) MATLAB version: 9.13.0 (R2022b), Natrick, Massachusetts: The MathWorks Inc. URL https://www.mathworks.com

Miyaoka, Osahito (1992) *Kita no Gengo: Ruikei to Rekishi*. Tokyo: Sanseidō.

Murawaki, Yugo (2019) Bayesian learning of latent representations of language structures. *Computational Linguistics*, 45(2): 199–228.

Nichols, Johanna (1994) The spread of language around the Pacific rim. *Evolutionary Anthropology: Issues, News, and Reviews*, 3(6): 206–215.

Ohsumi, Noboru (1989) *Tōkeiteki dēta kaiseki to sohutowea* [*Statistical data analysis and sotfware*]. Tokyo: Foundation for the Promotion of the Open University of Japan.

Ono, Yōhei (2020) How to handle "missing values" in linguistic typology: A pitfall in the statistical modeling approach. *Northern Language Studies*, 10: 61–82.

Ono, Yōhei and John Whitman (2016) Applying multiple correspondence analysis to non-word-order features reveals areal and genetic grouping in linguistic typology. *Research Memorandum of the Institute of Statistical Mathematics*, 1196.

Ono, Yōhei, Ryōzō Yoshino, Fumi Hayashi, and John Whitman (2017) A multiple correspondence analysis of the latent structure of features in linguistic typology (1):

A statistical reanalysis of Tsunoda, Ueda, and Itoh (1995a). *Mathematical Linguistics*, 31(3): 189–204.

Ono, Yōhei, Ryōzō Yoshino, Fumi Hayashi, and John Whitman (2018) A multiple correspondence analysis of the latent structure of features in linguistic typology (2): A statistical reanalysis of Tsunoda, Ueda, and Itoh (1995a). *Mathematical Linguistics,* 31(4): 261–280.

Rand, William M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66: 846–850.

R Core Team (2023) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sapir, Edward and Harry Hoijer (1967) *The phonology and morphology of the Navaholanguage*. Berkeley: University of California Press.

Simeon, George (1969) Hokkaido Ainu phonemics. *Journal of the American OrientalSociety*, 89(4): 751–757.

Svantesson, Jan-Olof (2003) Khalkha. In Janhunen, Juha (ed.), *The Mongolic languages*, 154–176. London: Routledge.

Sørensen, Thorvald (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, *Biologiske Skrifter*, 5: 1–34.

Takamura, Hiroya, Ryō Nagata, and Yoshifumi Kawasaki (2016) Discriminative analysis of linguistic features for typological study. In Nicoletta Calzolari et al. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 69–76. Retrieved from https://aclanthology.org/L16-1011/

Tibshirani, Robert, Walther Guenther, and Trevor Hastie (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B*, 63(2): 411–423.

Tsubaki, Hiroe (2011a) Tahenryō dēta to paneru dēta no soukan kouzou ni kansuru chuui to kokoromi [Some remarks and attempts on correlation structures on multivariate and panel data], *Collection of Technical Reports of the Second Workshop on Latent Dynamics*: 1–5. Retrieved from https://latent-dynamics.net/02/01_Tsubaki.pdf

Tsubaki, Hiroe (2011b) Tahenryō dēta to paneru dēta no soukan kouzou ni kansuru chuui to kokoromi [Some remarks and attempts on correlation structures on multivariate and panel data]. Presentation at Second Workshop on Latent Dynamics. Retrieved from https://latent-dynamics.net/02/01_Tsubaki.pptx.pdf

Tsubaki, Hiroe and Michiko Tsubaki (1997) Reconsideration on present statistical modeling from the viewpoints of graphical modeling. *Proceedings of 65st Annual Meeting of The Japan Statistical Society*, 256–257.

Ward, Joe, H., Jr. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58: 236–244.

Whitman, John and Yōhei Ono (2015) A preliminary list of features for a typological database of Northeast Asian (and North Pacific) languages. Presentation at Northeast Asia and the North Pacific as a linguistic area. Sapporo, Hokkaido University, 21 August 2015.

Whitman, John and Yōhei Ono (2017) Diachronic interpretation of word-order cohesion. In Mathieu, Éric and Robert Truswell (eds.), *Micro-change and Macro-change in Diachronic Syntax*, 43–60, Oxford: Oxford University Press.

Summary

As demonstrated by Ono (2020), previous "interdisciplinary" studies have misunderstood "missing values" in linguistic typology and imputed linguistically meaningless values on the missing values in the data. The main objectives of this paper are to propose exploratory statistical methods that enable researchers to approach areality and genealogy on languages by linguistic typology without imputing missing values.

We will propose new statistical analyses combining correspondence analysis (Benzécri et coll. 1973), correlation analysis (Tsubaki and Tsubaki 1997), graphical modeling (Lauritzen 1996), and clustering techniques that are based on three linguistic requirements in linguistic typology: "correction on frequencies," "profile view," and "language as variable."

The proposed methods resulted in not only outperforming previous studies on areal and genetic grouping in Africa, Eurasia, Europe or Indo-European family, and Austronesian and Austro-Asiatic but also detecting Circum-Pacific structure (Nichols 1994; Bickel and Nichols 2006) in languages. From the viewpoint of linguistics, our results showed that the word-order features have been masking areal and genetic groupings in linguistic typology, and word-order features related to Adjectives and Negation are needed in further investigations of linguistics. Since areal linguistics has mainly been focusing on some features or values that are not found in other areas or not necessarily included in WALS, our results opened up possibility that linguistic typology contributes to area linguistics by the common linguistic typological features or values.

Furthermore, the Circum-Pacific structure detected in this paper includes some Pacific Rim (Bickel and Nichols 2006). However, our classification result did not distinguish the coastal part of languages between the other parts in North and South America, the former of which strictly consist of Pacific Rim. Thus, improving the quality of the database in linguistic typology and developing statistical methodologies consistent with substantive viewpoints are still promising in addressing areality and genealogy, including Pacific Rim and North Pacific Rim (Miyaoka 1992) that cannot be traced back by another linguistic realm, including lexicostatistics.

(mathematical.humanities@hotmail.com)