Northern Language Studies 10: 231-254（Japan Association of NoLS, 2020）

Reconsideration of "Major Division" of Ainu Dialects: A Statistical Reanalysis of Asai (1974)

Yohei ONO
(Graduate Student at the Open University of Japan)

Keywords: Ainu, classification, Graph Theory, lexicostatistics, Spectral Clustering

1. Introduction

Humanities data are, in general, recorded using "various symbols" and these symbols are quantified based on substantive knowledge, consciously or implicitly. Therefore, researchers need to grasp the specific system of symbols and quantification in each field. These processes enable us to select or develop appropriate statistical methodologies and to understand the nature of the phenomenon. I began addressing this study motivated by the system on quantification, taking Asai's (1974) lexicostatistical research on Ainu dialects as an example.

Asai's (1974) study constitutes monumental research on Ainu linguistics, which established the classification of Ainu dialects based on Hattori and Chiri's (1960) lexicostatistical survey (Nos. 1-19 in Figure 1). Asai's (1974) contribution to Ainu linguistics may be summarized from the perspective of both the material and classification as follows.

First, from the perspective of materials in Ainu linguistics, Asai (1974) investigated original linguistic materials on the Obihiro, Kushiro, Asahikawa, and Chitose dialects (Nos. 8, 9, 11, and 21 in Figure 1) and gathered the data for the North Kuril dialect (No. 20 in Figure 1) with reference to Torii (1903), Murayama (1971), and Pinart's vocabulary manuscript (Asai 1974: Appendix). Therefore, his documentation has made an invaluable contribution to Ainu linguistics in the area.

Furthermore, as shown in Table 1, Asai (1974: 92; Table 1) demonstrates the cognacy judgments on 110 words among the 21 Ainu dialects. Since no other documents exist where Ainu linguists demonstrate the cognacy judgments among the Hokkaido Ainu dialects (Nos. 1-13, and 21 in Figure 1), North Kuril Ainu dialect (No. 20 in Figure 1), and Sakhalin Ainu dialects (Nos. 14-19 in Figure 1), Asai's (1974) study is still a landmark in Ainu dialectology.

Second, from the perspective of the classification of the Ainu dialects, Asai (1974) applied a cluster analysis to his data matrix (Asai 1974: 92; Table 1) and illustrated the classification of the Ainu dialects that still has a great influence on current Ainu linguistics.

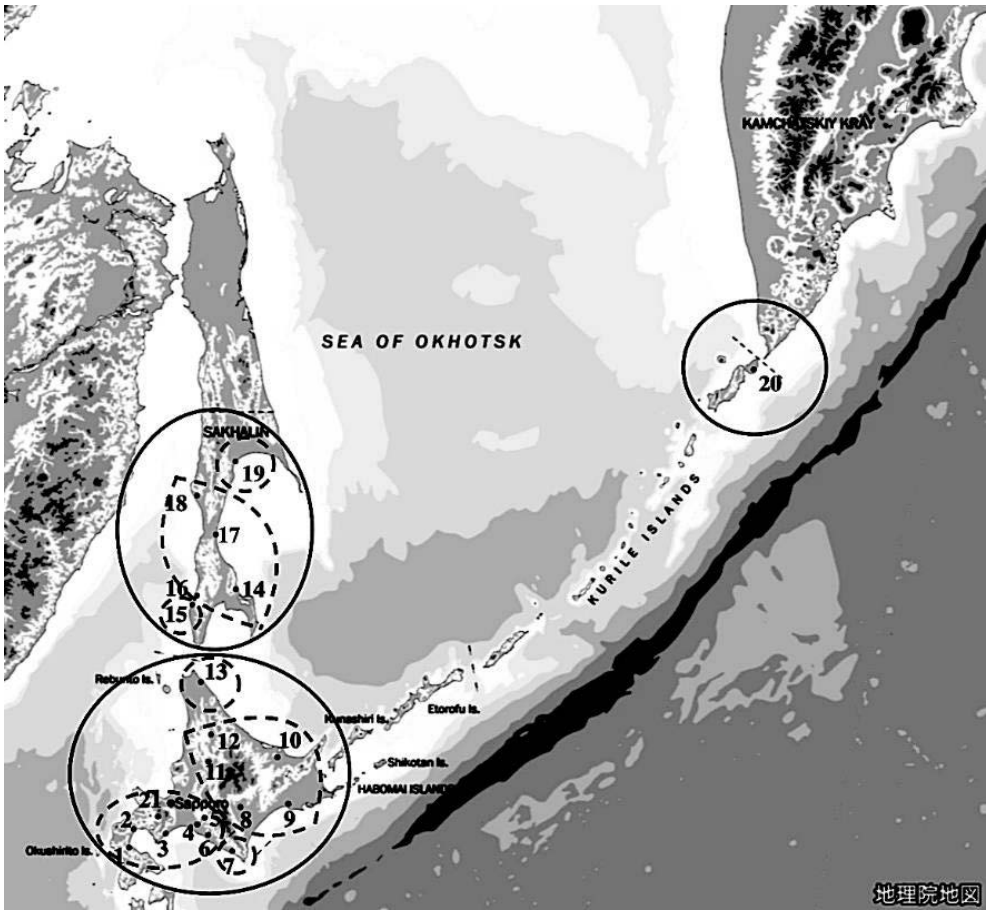

Figure 1. Map of a section of the region where the Ainu language is or was spoken (Geospatial Information Authority of Japan 2019), edited by the author. 1: Yakumo, 2: Oshamambe, 3: Horobetsu, 4: Biratori, 5: Nukibetsu, 6: Niikappu, 7: Samani, 8: Obihiro, 9: Kushiro, 10: Bihoro, 11: Asahikawa, 12: Nayoro, 13: Soya, 14: Ochiho, 15: Tarantomari, 16: Maoka, 17: Shiraura, 18: Raichishka, 19: Nairo, 20: North Kuril (Shumushu), 21: Chitose. Note that "Major division" is marked by circles and "Minor Division" by dotted circles in Asai (1974: 100). "Minor division" in Asai (1974: 100) is as follows: North Hokkaido: 13; East Hokkaido: 8, 9, 10, 11, and 12; Central South Hokkaido: 1, 2, 3, 4, 5, 6, and 21; Eastern Hokkaido: 7; North Sakhalin: 19; Central Sakhalin: 14, 16, 17, and 18; South Sakhalin: 15; North Kuril: 20.

The significant findings of his study demonstrated that the Ainu dialects can be classified into three groups: the Hokkaido Ainu dialects (Nos. 1-13, and 21 in Figure 1), North Kuril Ainu dialect (No. 20 in Figure 1), and Sakhalin Ainu dialects (Nos. 14-19 in Figure 1). Although Kindaichi (1932) already noted these classifications of Ainu dialects, Asai's (1974) main results comprised the establishment of these classifications

with statistical methods, that is, with “objective criteria.”

Asai (1974: 100) called the classification that distinguished the Hokkaido, North Kuril, and Sakhalin Ainu dialect groups as the “Major Division” and the sub-classification of the “Major Division” as the “Minor Division”, as demonstrated in Figure 1.

Table 1. The data on cognacy judgments on 110 words among the 21 Ainu dialects from Asai (1974: 92; Table 1). In the following tables, the row number corresponds to each dialect in the first column.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1_Yakumo	110	104	102	97	96	95	83	88	88	83	94	89	77	40	52	43	39	42	43	54	94
2_Oshamambe	104	110	101	95	94	93	80	84	84	78	92	84	77	37	50	39	36	41	42	50	95
3_Horobetsu	102	101	110	104	98	98	83	91	90	85	96	93	76	41	56	45	41	44	46	59	99
4_Biratori	97	95	104	110	103	105	79	87	86	81	92	89	73	42	58	47	42	45	48	59	104
5_Nukibetsu	96	94	98	103	110	98	75	81	81	74	87	83	72	38	51	41	36	40	41	54	101
6_Niikappu	95	93	98	105	98	110	73	84	84	78	89	83	70	40	54	43	39	42	45	56	101
7_Samani	83	80	83	79	75	73	110	94	92	84	85	81	74	33	49	36	33	33	38	52	80
8_Obihiro	88	84	91	87	81	84	94	110	106	100	95	93	79	38	54	43	38	31	46	59	84
9_Kushiro	88	84	90	86	81	84	92	106	110	101	98	98	83	38	54	42	38	41	45	63	85
10_Bihoro	83	78	85	81	74	78	84	100	101	110	92	91	78	36	52	42	36	39	43	58	79
11_Asahikawa	94	92	96	92	87	89	85	95	98	92	110	102	82	43	58	47	40	45	47	60	93
12_Nayoro	89	84	93	89	83	83	81	93	98	91	102	110	84	44	59	51	42	48	49	61	87
13_Soya	77	77	76	73	72	70	74	79	83	78	82	84	110	47	65	56	48	55	52	51	72
14_Ochiho	40	37	41	42	38	40	33	38	38	36	43	44	47	110	66	90	91	84	60	29	40
15_Tarantomari	52	50	56	58	51	54	49	54	54	52	58	59	65	66	110	78	67	67	74	44	55
16_Maoka	43	39	45	47	41	43	36	43	42	42	47	51	56	90	78	110	92	87	68	32	46
17_Shiraura	39	36	41	42	36	39	33	38	38	36	40	42	48	91	67	92	110	93	68	27	40
18_Raichishka	42	41	44	45	40	42	33	31	41	39	45	48	55	84	67	87	93	110	62	33	41
19_Nairo	43	42	46	48	41	45	38	46	45	43	47	49	52	60	74	68	68	62	110	38	39
20_North_Kuril	54	50	59	59	54	56	52	59	63	58	60	61	51	29	44	32	27	33	38	110	54
21_Chitose	94	95	99	104	101	101	80	84	85	79	93	87	72	40	55	46	40	41	39	54	110

Note that Asai (1974: 92; Table 1) focuses on whether there is at least one common word form of the word among the Ainu dialects or not, as described in the “relation index” in Asai (1974: 61-62).

However, as a statistician, I was confronted with two questions on Asai’s (1974) main result, from the viewpoint of statistical methodologies. First, there were other clustering methods considered as able to produce a more reliable classification than “Large method” and “Small method” at the time Asai (1974) was published: the unweighted pair group method with arithmetic mean (Sokal and Michener 1958; often abbreviated as UPGMA) also known as the “group-average method” and Ward’s minimum variance method (Ward 1963)[1]. However, without any explanation, Asai (1974) did not choose these clustering methods.

Second, clustering methods, such as the “Large method” and “Small method”

[1] Everitt (1979: 173) writes that, although no single method is best for all situations, the mathematically respectable nearest-neighbor method (i.e., the same approach used in Asai’s [1974] “Large method”) is, in most cases, the least successful for the data used. Moreover, he also writes that group-average clustering and Ward’s minimum variance method perform fairly well overall.

applied to Asai's (1974: 92; Table 1) data matrix, are considered appropriate for analyzing the partial information on the cognacy judgments on 110 words among the 21 Ainu dialects.

Therefore, I started conducting this study through the consideration of these questions. As demonstrated in Section 2, Asai (1974: 92; Table 1) contains a "special data structure", under which it is meaningless, from a linguistic perspective, for researchers to subtract similarity data in Asai (1974) and utilize these values as "dissimilarity." Consequently, among clustering methods, only "Large method" and "Small method" could be applied to the "special data structure" in Asai (1974: 92; Table 1) at the time Asai (1974) was published[2].

This study proposes an alternative statistical method, Spectral Clustering, popularized in the realm of machine learning by Shi and Malik (2000) and Ng, Jordan, and Weiss (2002) after a quarter of a century.

The purpose of Spectral Clustering is to discover the division of the objects that minimizes the ratio of the sum of the similarity in the groups to the sum of the similarity among the groups; in other words, the division aims to maximize the sum of the similarities in the groups and minimize the sum of the similarities among the groups.

As demonstrated in Section 2, Spectral Clustering can be applied to "some special algebraic structure" in Asai (1974: 92; Table 1) and uncover the underlying structure in the data more effectively than Asai's (1974) "Large method" and "Small method."

Therefore, the main objective of this study is to reexamine the classifications of Ainu dialects, focusing on "Major Division" and "Minor Division" that Asai (1974) "established" with objective criteria, through the application of a more appropriate statistical method, Spectral Clustering.

The main results demonstrate that the "Major Division" of Ainu dialects, which classifies them into three groups—Hokkaido Ainu dialects (Nos. 1-13, and 21 in Figure 1), North Kuril Ainu dialect (No. 20 in Figure 1), and Sakhalin Ainu dialects (Nos. 14-19 in Figure 1)—needs to be reconsidered, so far as researchers recognize the classification of Ainu dialects based on Asai's (1974) clustering results.

Furthermore, the results of Spectral Clustering illustrate that the 21 Ainu dialects can be classified into two groups: Hokkaido Ainu dialects (Nos. 1-13, 20, and 21 in Figure 1) and Sakhalin Ainu dialects (Nos. 14-19 in Figure 1). Notably, the North Kuril

[2] Notably, Asai (1974: 62-63) introduces other clustering methods called the "algebraic mean method" and "geometric mean method", the former of which corresponds to the "group-average method." However, Asai (1974) did not adopt these clustering methods in the analysis without offering any explanation. As discussed in Section 2, I demonstrate that these two clustering methods apply division and *n*th root to the "special data structure" in Asai (1974: 92; Table 1), and it is questionable, from a linguistic perspective, that researchers use these fractional or irrational numbers as "similarity" among the dialects. These facts indicate that Asai himself noticed the particularity of the data structure in Asai (1974: 92; Table 1).

Ainu dialect (No. 20 in Figure 1) belongs to the northeastern Hokkaido Ainu dialect group (Nos. 7-13 in Figure 1) in the strongest structure and to southwestern Hokkaido Ainu dialect group (Nos. 1-6, and 21 in Figure 1) in the second strongest structure among the Hokkaido Ainu dialects (Nos. 1-13, 20, and 21 in Figure 1). Moreover, the sub-classification of the Ainu dialects obtained by Spectral Clustering exhibits a structure different from Asai’s (1974) “Minor Division” in several points.

Therefore, these results suggest a need to reconsider the classification of Ainu dialects, including the linguistic research based on Asai’s (1974) clustering results and the systematic review of the cognacy judgments in Asai (1974: 92; Table 1) in current Ainu linguistics.

The remainder of this paper is organized as follows: Section 2 discusses the basic data property in Asai (1974: 92; Table 1) and the statistical methodologies considered to be more appropriate for the analyses of Asai (1974: 92; Table 1), from the viewpoints of both linguistics and statistics. The discussions lead me to apply Spectral Clustering to the data in question. Furthermore, a criterion for evaluating the three clustering methods (i.e., “Large method”, “Small method”, and Spectral Clustering) is also introduced.

Section 3 applies Spectral Clustering to several different similarity matrices that “Large method” and “Small method” cannot identify as the result of the classification, and illustrates that Spectral Clustering can identify these matrices. These results demonstrate that Spectral Clustering is more appropriate for the analyses of Asai (1974: 92; Table 1) in terms of “identifiability”—the evaluation criterion introduced in Section 2. Therefore, I consider the classification of Ainu dialects based on the results of Spectral Clustering applied to Asai (1974: 92; Table 1).

Section 4 discusses the significance of this study. The choices of “similarity” or “dissimilarity” index in linguistic data and its relation to applicable statistical methods are discussed from a statistical viewpoint. Furthermore, I propose a classification of Ainu dialects, which places the dialects into the Hokkaido Ainu dialect group, including the North Kuril dialect, and the Sakhalin Ainu dialect group, from both linguistic and philological points of view[3].

The main result of this study is that the “Major Division” of Ainu dialects is not supported from a mathematical and statistical point of view. This indicates that through his “Major Division” (and “Minor Division”), Asai (1974) may not have intended to “establish” but rather has attempted to apply clustering methods as a reference for

[3] I note that Murayama (1971: 127-131) has already noted that the North Kuril dialect is closely related to the north areas of Hokkaido Ainu dialects and has a few common features with the Sakhalin Ainu dialects based on a lexical analysis. Thus, the classification found in this study coincides with the linguistic and philological research conducted before Asai (1974). Moreover, Satō and Bugaeva (2019: 78-86) have also illustrated that some items in unpublished 19th Kuril Ainu documents cluster with Southern Hokkaido (Saru/Chitose). Thus, the statistical findings in this study coincide with these two philological analyses.

further linguistic research; however, these classifications resulted in a great influence on current Ainu linguistics, beyond the purpose of Asai himself.

2. Materials and Methods

Section 2 focuses on the data in Asai's Table 1 (1974: 92) and discusses statistical methodologies that can analyze Asai's Table 1 (1974: 92) more appropriately from both a linguistics and statistics viewpoint. Furthermore, I introduce "identifiability" as the evaluation criterion of the clustering methods in Section 2.3. Spectral Clustering is introduced in Section 2.4 as an alternative method to Asai's (1974) "Large method" and "Small method."

Section 2.1 discusses the basic property in Asai (1974: 92; Table 1) with specific materials and demonstrates that it contains only the information on "similarity" among Ainu dialects. Thus, Asai (1974: 92; Table 1) does not necessarily preserve the information on "dissimilarity" among Ainu dialects.

Section 2.2 deals with the special data structure in Asai (1974: 92; Table 1). Since Asai (1974: 92; Table 1) contains only the information on similarity among the Ainu dialects, it is meaningless, from a linguistic perspective, for researchers to subtract "similarities" from Asai (1974: 92; Table 1) and define these values as "dissimilarity" among the dialects. Therefore, this violates the assumptions of classical statistical methods, including those of (metric or non-metric) Multidimensional Scaling. Moreover, I illustrate that clustering methods other than "Large method" and "Small method" (e.g., "algebraic mean method" and "geometric mean method") apply division and *n*th root to Asai (1974: 92; Table 1). Thus, it is questionable from a linguistic perspective for researchers to compare these fractional or irrational numbers as "similarity" among the dialects.

Section 2.3 illustrates that "Large method" and "Small method" produce the same classification with Asai (1974: 92; Table 1) based on several different "similarity" matrices, respectively. The statistical properties of "Large method" and "Small method" correspond to the problem of "identifiability" in terms of statistics: the two methods cannot identify several different "similarity" matrices as the result of the classification.

Since Asai (1974: 46; Footnote 1) states[4], "*Here we mean a classification that presupposes any kind of hierarchical structure or a grouping as a most feasible method*", the lack of "identifiability" of "Large method" and "Small method" indicates that these clustering methods are insufficient for classifying Ainu dialects based on Asai (1974: 92; Table 1).

Furthermore, I state the criterion for evaluating clustering methods other than "Large method" and "Small method." Since "identifiability" for different similarity

[4] In quotations, unless italicized, the English translation from Japanese is by the author.

matrices is considered a desirable property, I adopt the criterion of whether a clustering method can identify, as the result of the classification, several different similarity matrices that "Large method" and "Small method" cannot identify in Section 2.3.

Section 2.4 introduces Spectral Clustering as an alternative clustering method. Since Spectral Clustering does not subtract data or define their values as "dissimilarities" and only utilize addition and comparison, it is meaningful, from a linguistic perspective, that researchers apply Spectral Clustering to Asai (1974: 92; Table 1). The background on Spectral Clustering is also introduced.

2.1 Materials from a Linguistic Perspective

Section 2.1 discusses the basic property of Asai (1974: 92; Table 1) from linguistic perspectives. Asai (1974: 67) explained that "*obviously similar or nearly similar Ainu forms are also put in parentheses, but considerably many forms, which may be easily taken as cognates, are not placed in parentheses*." However, he did not clearly state whether Asai (1974: 92; Table 1) considered the word forms put in parentheses as a one-word form or not. For the sake of convenience in the description, I tentatively categorize the word forms in parentheses as a one-word form.

According to Asai's own linguistic knowledge (Asai 1974: 67), the word forms on 'I' are as follows; (kuani, kani)(1-12, 17-21); (cokaj, cookaj, ciokaj)(13, 16, 17, 19, 21); (anoka, anokaj)(14-16, 18, 19)[5].

However, the significant question soon arises regarding how to "appropriately" calculate similarity from the data and what operation (e.g., addition, subtraction, multiplication, division, or *n*th root) is allowed to be performed on the calculated similarities[6].

We observe that the number of word forms in certain dialects severely biases the data on similarity among the dialects. For example, the Nairo dialect has three word forms: (kuani, kani), (cokaj, cookaj, ciokaj), and (anoka, anokaj), and tends to share word forms with other dialects; in contrast, the Yakumo dialect has one-word form: (kuani, kani). Therefore, the calculation of similarity among the Ainu dialects suffers from this critical phenomenon and leads to a "special data structure", as mentioned in Section 1. Thus, Asai (1974: 92; Table 1) was derived through an alternative approach that focused on whether there exists at least one common word form of one word among the dialects, as described by the definition of "relation index" in Asai (1974: 61-62).

---

[5] The numbers in parentheses following each word form correspond to the places in Figure 1.

[6] There are various ways to calculate "similarity" or "dissimilarity" among the dialects (or the languages) from lexicostatistical data, and the problem has been argued about heretofore. Asai (1974: 49-57) explained his position by stating that he considered the remaining words in the basic word list between two dialects as the linguistical distance; this standpoint is closely related to "Lexicostatistic Dating" or "Glottochronology", established by R. B. Lees and M. Swadesh. Therefore, Asai (1974: 92; Table 1) utilizes the information on "similarity" among the Ainu dialects.

Table 2. The presence of common word forms on the word 'I' among the Ainu dialects.

Dialect	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1_Yakumo	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
2_Oshamambe	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
3_Horobetsu	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
4_Biratori	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
5_Nukibetsu	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
6_Niikappu	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
7_Samani	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
8_Obihiro	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
9_Kushiro	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
10_Bihoro	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
11_Asahikawa	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
12_Nayoro	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
13_Soya	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	1
14_Ochiho	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0
15_Tarantomari	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0
16_Maoka	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	1
17_Shiraura	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1
18_Raichishka	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
19_Nairo	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20_North_Kuril	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
21_Chitose	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1

Table 2 demonstrates similarity data among the word forms on 'I' using the "relation index." Asai (1974: 61-62) explained that "*we assume that the value of the relation index is 1, if there is at least one similar or the same form in both of any two given dialects (P_i, P_j) and if there is no common form in the two given dialects, the relation index of the two dialects equals a value of 0*".

For example, the Maoka dialect and the Nairo dialect share two word forms (i.e., [cokaj, cookaj, ciokaj] and [anoka, anokaj]) but the similarity data between the two dialects demonstrates 1 in the definition of "relation index".

Consequently, Asai (1974: 92; Table 1) comprises the summation of 110 words using the "relation index.[7]" Since this index focuses on whether there remains at least one word form of a word among the dialects, Asai (1974: 92; Table 1) does not contain any information on how many word forms the two given dialects share or do not share.

Therefore, the facts demonstrate that Asai (1974: 92; Table 1) contains only information about the similarity among the Ainu dialects; in other words, researchers cannot construct any negative numbers that represent the opposites of the positive similarity data in Asai (1974). Again, note that the "relation index" states "*we assume that the value of the relation index is 1, if there is <u>at least</u> one similar or the same form*

[7] There still remains an unsolved problem in Ainu linguistics: whether researchers can specify the 110 words that Asai (1974: 92; Table 1) utilized, but did not demonstrate, among the linguistic data in Asai (1974: 67-93). Ono (2020, to appear) concluded that the description in Asai (1974) is insufficient for specifying 110 words from a statistical viewpoint. Therefore, I leave this problem to Ono (2020, to appear) and assume the data in Asai (1974: 92; Table 1) in the subsequent analysis.

in both of any two given dialects (P_i, P_j) and if there is no common form in the two given dialects, the relation index of the two dialects equals a value of 0" (Asai 1974: 61-62).

Therefore, -1, a negative number, requires the condition that, if researchers add it to a +1, that is, "*there is <u>at least</u> one similar or the same form in both of any two given dialects (P_i, P_j)*", this leads to a value of 0, implying that "*there is no common form in the two given dialects, the relation index of the two dialects equals a value of 0*".

However, the definition of 1 in Asai (1974: 61-62) is, literally, the negation of the definition of 0 in Asai (1974: 61-62), and the definition of 0 in Asai (1974: 61-62) is, literally, the negation of the definition of 1 in Asai (1974: 61-62). Furthermore, both the definition of 1 and 0 cover all possibilities about the judgment of one word.

Therefore, the logical interpretation in Asai (1974: 61-62) precludes the possibility of defining -1 without contradiction. Thus, it is meaningless, from a linguistic perspective, that researchers subtract values in Asai (1974: 92; Table 1) and define its values as dissimilarities in any statistical calculation.

As a result, the basic property of Asai (1974: 92; Table 1) violates the assumptions of classical statistical methods, as discussed in Section 2.3.

2.2 Materials from a Statistical Perspective

This section discusses Asai (1974: 92; Table 1) from a statistical perspective. Although researchers can "perform" a subtraction operation and utilize Asai (1974: 92; Table 1) as "dissimilarities", these values are meaningless from a linguistic perspective[8]. These results lead to the violation of the assumptions of most statistical analyses applied to linguistic data. For example, statistical methods involving the inner scalar product (e.g., metric Multidimensional Scaling and Correspondence Analysis) assume that the subtraction of data is meaningful from a substantive knowledge perspective (i.e., linguistic knowledge in this case).

For example, metric Multidimensional Scaling utilizes an inner scalar product

[8] Here, I note that Asai (1974: 92; Table 1) may be formularized as a commutative monoid (or abelian monoid) on addition in algebra (cf. Gondran and Minoux 2008). A commutative monoid has the mathematical properties of a single associative and commutative binary operation, a neutral element, an order-unit, and an algebraic preorder relation, but not an inverse element. The meaning of these five properties is as follows: (1) a single associative and commutative binary operation: addition (+) (researchers are allowed to add the data); (2) a neutral element: the existence of a 0, or, Asai (1974: 61-62) defines the non-existence of a cognate word form of one word in two dialects as a 0; (3) an order-unit: the existence of a 1, or, Asai (1974: 61-62) defines the existence of at least one common word form between two dialects as a 1; (4) preorder relation (≦): it is meaningful, from a linguistic perspective, that researchers add the values of Asai (1974: 92; Table 1) and compare them; (5) non-existence of an inverse element: the non-existence of subtraction, or, it is meaningless, from a linguistic perspective, that researchers subtract the values of Asai (1974: 92; Table 1), compare, and define them as the "dissimilarities" between the dialects. These mathematical perspectives might potentially contribute to the development of linguistics.

model under the assumption of the equation[9]:

$$d_{jk}^2 = s_{jj} + s_{kk} - 2s_{jk}.$$

Notably, the third term on the right side in the equation requires the subtraction of similarity data (or the inverse element of similarity data). Therefore, the equation clearly violates the assumption in Asai (1974: 92; Table 1). Since an inner scalar product model is utilized in most of the statistical techniques for the visualization of data structure, this fact indicates the difficulties that lead researchers to directly follow Asai (1974: 92; Table 1) as a basis for their analyses.

Moreover, even the application of non-metric Multidimensional Scaling (Kruskal 1964a; 1964b) is meaningless from a linguistic perspective. The assumption of non-metric Multidimensional Scaling that similarity data preserve the ordering in dissimilarity data is represented as follows:

$$s_{ik} < s_{jk} \Leftrightarrow d_{ik} > d_{jk}$$

As demonstrated in Section 2.1, Asai (1974: 92; Table 1) contains only information on the "similarity" among the Ainu dialects and does not guarantee information on the "dissimilarity" among the Ainu dialects.

Since metric Multidimensional Scaling (e.g., Black 1973) and non-metric Multidimensional Scaling (e.g., Dyen, Kruskal, and Black 1992) have been applied to lexicostatistical data similar to Asai (1974: 92; Table 1), this suggests a need to reconsider previous lexicostatistical research utilizing statistical methods based on the two assumptions explained above.

Next, I discuss the question of whether clustering methods other than "Large method" and "Small method" (e.g., "algebraic mean method", "geometric mean method", and Ward's minimum variance method) can be applied to Asai (1974: 92; Table 1).

These clustering methods belong to a kind of "hierarchical cluster analysis" in statistics. Since hierarchical cluster analysis agglomerates objects based on its own criterion as a hierarchy, various characteristics in different hierarchical clustering methods consist of their own criterion defining the similarity or dissimilarity between groups of objects.

For example, hierarchical clustering methods first cluster two objects that show the largest similarity in a similarity matrix like Table 1. Then, a significant question arises

[9] The meaning of the mathematical symbols used hereafter is as follows: d_{jk}: dissimilarity between dialect *j* and dialect *k*; s_{jk}: similarity between dialect *j* and dialect *k*; $s_{i(j\cup k)}$: similarity between dialect *i* and the group of dialect *j* and dialect *k*; *max()*: a function that chooses the term with the maximum value (i.e., among the similarities of the dialects in Asai [1974: 92; Table 1]); *min()*: a function that chooses the term with the minimum value (i.e., among the similarities of the dialects in Asai [1974: 92; Table 1]).

as to how to define the similarity between the two objects and the other objects, $s_{i(j\cup k)}$[10].

First, Ward's minimum variance method forms a hierarchy of clusters based on the dissimilarity among the data. Therefore, it is meaningless, from a linguistic perspective, that a researcher applies Ward's minimum variance method to Asai (1974: 92; Table 1).
Next, the "algebraic mean method" defines the similarity between dialect *i* and the group of dialect *j* and dialect *k* (e.g., the two objects that show the largest similarity in a similarity matrix like Table 1) as follows:

$$s_{i(j\cup k)} = \frac{1}{2}\left(s_{ij} + s_{ik}\right)$$

Furthermore, the "geometric mean method" defines the similarity between dialect *i* and the group of dialect *j* and dialect *k* as follows:

$$s_{i(j\cup k)} = \sqrt{s_{ij}s_{ik}}$$

We observe that "algebraic mean method" and "geometric mean method" can utilize a fractional and an irrational number, respectively. Since the similarity data in Asai (1974: 92; Table 1) consist of non-negative "integers" as demonstrated in Section 2.1, it is questionable, from a linguistic perspective, that researchers apply and compare these fractional or irrational numbers as a "similarity" among the dialects.

Note that Asai (1974: 62-63) introduces the "algebraic mean method" and "geometric mean method" but does not utilize them in the analysis, without any explanation. This fact indicates that Asai himself noticed the particularity of his data.

However, "Large method" and "Small method" can be applied to Asai (1974: 92; Table 1). "Large method" defines the similarity between dialect *i* and the group of dialect *j* and dialect *k* as follows:

$$s_{i(j\cup k)} = max(s_{ij}, s_{ik})$$

"Small method" defines the similarity between dialect *i* and the group of dialect *j* and dialect *k* as follows:

$$s_{i(j\cup k)} = min(s_{ij}, s_{ik})$$

Since *max()* in "Large method" and *min()* in "Small method" are feasible with comparison, I assume that among the well-known clustering methods, only "Large method" and "Small method" could be applied to Asai (1974: 92; Table 1) at the time Asai (1974) was published. However, Section 2.3 illustrates that "Large method" and "Small method" are also insufficient for a statistical analysis of Asai (1974: 92; Table 1) based on the concept of "identifiability."

[10] Note that the similarity between dialect group A and a group consisting of dialect group B and dialect group C can be defined using the recursive application of the definition of $s_{i(j\cup k)}$.

2.3 Methodological Problem in "Large method" and "Small method"

This section demonstrates the shortcomings of "Large method" and "Small method" based on the concept of "identifiability." Figure 2 shows the clustering results of "Large method" and "Small method" when applied to Table 1[11,12].

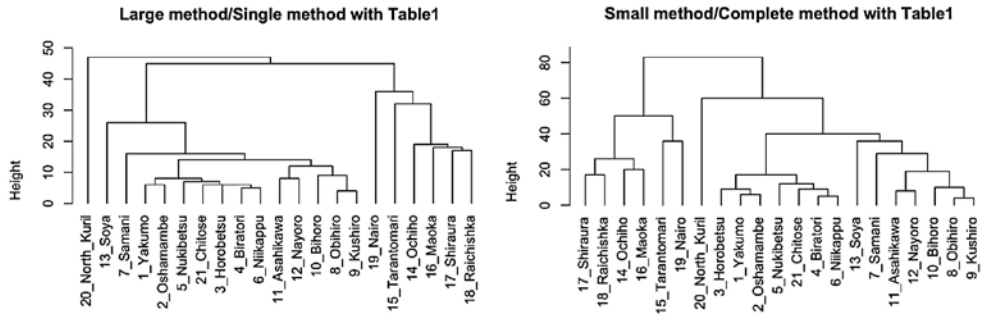

Figure 2. The clustering results of Table 1. Left: Large method. Right: Small method

However, the similarities among the North Kuril dialect (No. 20 in Figure 1) and the Hokkaido Ainu dialect group (Nos. 1-13 and 21 in Figure 1) do not affect the clustering result (i.e., the dendrogram), as I sort these values (i.e., grey cells in Table 1) arbitrarily. Thus, I create Data 1, 2, and 3 from Table 1 by sorting the values among the North Kuril dialect and the Hokkaido Ainu dialect group (No. 20 in Figure 1 and Nos. 1-13, and 21 in Figure 1) as follows:

Table 1: 54, 50, 59, 59, 54, 56, 52, 59, 63, 58, 60, 61, 51, 54;
Data 1: 63, 61, 60, 54, 54, 54, 52, 51, 50, 56, 59, 59, 59, 58;
Data 2: 50, 51, 52, 56, 58, 59, 59, 54, 54, 59, 60, 61, 63, 54;
Data 3: 54, 54, 58, 59, 60, 61, 63, 59, 59, 56, 52, 51, 50, 54.

Data 1, 2, and 3 are generated to locate the North Kuril dialect (No. 20 in Figure 1) near a part of the Hokkaido Ainu dialects (Nos. 1-3 and 11-13 in Figure 1), near the North Hokkaido dialect (No. 13 in Figure 1), and near the Eastern Hokkaido dialect (No. 7 in Figure 1), respectively. Due to space restrictions, the clustering results of "Large method" and "Small method" are omitted. I observed that "Large method" applied to

[11] The clustering procedures of "Large method" and "Small method" correspond to the single method (or nearest-neighbor method) and the complete linkage method (Sørensen 1948), respectively. Therefore, for the sake of convenience, I applies the single method and complete linkage method to the data obtained by subtracting the similarity data from the number of words (i.e., 110 words in Asai [1974: 92; Table 1]) in the subsequent analyses.

[12] There is the problem with similarity ties in both the "Large method" (or single method) and "Small method" (or complete linkage method). In the subsequent analyses, "Large method" or single method implemented in R (R Core Team 2018) represents similarity ties automatically but "Small method" or complete linkage method corresponds to the result of "Small method B" in Asai (1974: 94).

Table 1 and Data 1-3 yields identical clustering results for the Ainu dialects, and “Small method” applied to Table 1 and Data 1-3 also produces identical clustering results for the Ainu dialects. These results demonstrate that “Large method” and “Small method” cannot identify Table 1 and Data 1-3 as the clustering results.

Furthermore, “Large method” produces the identical clustering as the left on Figure 2 in the case that the maximal value among the North Kuril dialect and the Hokkaido Ainu dialect group (No. 20 in Figure 1 and Nos. 1-13, and 21 in Figure 1) is equal to those in Table 1 (i.e., 63), and the minimal value among these dialects is equal to those in Table 1 (i.e., 50). Also, “Small method” produces the identical clustering as the right on Figure 2 in the same case. Thus, I create Data 4, 5, 6, and 7 from Table 1 by changing the values among the North Kuril dialect and the Hokkaido Ainu dialect group with the conditions above as follows:

Data 4: 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 63;
Data 5: 63, 63, 63, 63, 63, 63, 63, 50, 50, 50, 50, 50, 50, 50;
Data 6: 50, 63, 50, 63, 50, 63, 50, 63, 50, 63, 50, 63, 50, 63;
Data 7: 63, 50, 63, 50, 63, 50, 63, 50, 63, 50, 63, 50, 63, 50.

As discussed in Section 2.2, “Large method” defines the similarity between dialect *i* and the group of dialect *j* and dialect *k* as follows:

$$s_{i(j\cup k)} = max(s_{ij}, s_{ik})$$

“Small method” defines the similarity between dialect *i* and the group of dialect *j* and dialect *k* as follows:

$$s_{i(j\cup k)} = min(s_{ij}, s_{ik})$$

Since *max()* in “Large method” utilize only the information on the largest value and *min()* in “Small method” utilize only the information on the smallest value among the objects to agglomerate by constructing a hierarchical tree on the objects as Figure 2, “Large method” tends to ignore the values (or information) under the largest value and “Small method” tends to ignore those above the smallest value among the objects to agglomerate. Therefore, “Large method” and “Small method” often result in not reflecting the values between the smallest value and the largest value among the objects to agglomerate (i.e., among the North Kuril dialect and Hokkaido Ainu dialects).

Thus, “Large method” applied to Table 1 and Data 1-7 yields identical clustering results for the Ainu dialects, and “Small method” applied to Table 1 and Data 1-7 also produces identical clustering results for the Ainu dialects.

Since the resulting classification should reflect the information among the objects (i.e., the Ainu dialects in this case) as exactly as possible from the linguistic perspective, the mathematical property is not admissible from the substantive knowledge. Note that Asai (1974) and this paper do not focus on genealogy or phylogeny with lexicostatistical data in Ainu, but on classification in Ainu dialects.

Moreover, the purpose of Asai’s (1974) cluster analysis is to discover the

classification of Ainu dialects with the most feasible method. Then, the property of "Large method" and "Small method" explained above is considered to be undesirable for this purpose; in other words, an alternative clustering method should identify Table 1 and Data 1-7 as the clustering results. Thus, I define "identifiability" as the property that clustering methods can distinguish among different similarity matrices as the result and adopt "identifiability" as the evaluation criterion for clustering methods.

Section 2.4 introduces an alternative method, Spectral Clustering, which can be applied to similarity data with the mathematical structure in Asai (1974: 92; Table 1) and avoid the disadvantages of "Large method" and "Small method."

2.4 Spectral Clustering

This section introduces Spectral Clustering and its advantages. First, as noted in Section 1, the purpose of Spectral Clustering is to discover the division of objects that minimizes the ratio of the sum of the similarity in the groups to the sum of the similarity among the groups; in other words, the division aims to maximize the sum of the similarity in the groups and minimize the sum of the similarity among the groups.

In the case of 2 divisions, the objective function of the MinmaxCut algorithm (Ding et al. 2001; abbreviated as Mcut algorithm hereafter) that I adopt in this study is[13]:

$$MCUT(A,B) = \frac{CUT(A,B)}{W(A)} + \frac{CUT(A,B)}{W(B)}$$

The advantages of Spectral Clustering are summarized as follows: First, $CUT(A,B)$, $W(A)$, and $W(B)$ can be calculated by only adding similarity data, which is applicable to Asai (1974: 92; Table 1) as demonstrated in Sections 2.1 and 2.2; in other words, $CUT(A,B)$, $W(A)$, and $W(B)$ can be applied to data that contain only similarity information[14,15]. Second, $CUT(A,B)$, $W(A)$, and $W(B)$ utilize the entire

---

[13] $CUT(A,B)$: Sum of the similarities between group A and group B; $W(A)$: sum of the similarities in group A; $W(B)$: sum of the similarities in group B.

[14] It is noted that researchers can optimize $CUT(A,B)$ with addition and comparison admissible for commutative monoid. However, as von Luxburg (2007: 401) states, the intuitive objective function does not produce a satisfactory classification in practice. Since a smaller number of objects in the partition reduces $CUT(A,B)$, the solution tends to simply separate one object from the others, which is undesirable for the classification of dialects in this paper.

[15] $MCUT(A,B)$ utilizes division to summations of data (i.e., $CUT(A,B)$, $W(A)$, and $W(B)$) and addition to their quotients. Hence, some readers may pose question on these operations as explained in Section 2. $MCUT(A,B)$ can reformulate as follows:

$$MCUT(A,B) = CUT(A,A^c)/W(A) + CUT(B,B^c)/W(B)$$

Note that A^c is the complement set of A and B^c is the complement set of B. Here the first term of the right equation measures the ratio of the sum of the similarity in A to the sum of the similarity between A and the others, which corresponds to the goodness of A as partition intuitively, and the second term of the right equation also measures the ratio of the sum of the similarity in B to the sum of the similarity between B and the others, which also corresponds to the goodness of B as partition

information of the similarity matrix, in contrast to "Large method" or "Small method", which use only the maximal or minimal values in the similarity matrix. Third, when a researcher needs more than 3 divisions, the Mcut algorithm searches A and B for a subdivision. This property of subdivision of the Mcut algorithm yields a hierarchy of objects (e.g., the hierarchy of the dialects in our case) that corresponds to the notion of hierarchical clustering methods in this study and is desirable for the classification of dialects[16]. I utilize the Mcut algorithm implemented with R (R Core Team 2018) in Shinnou (2007: 132-133) in the subsequent analyses.

3. Results

Figures 3-4 present the results of the Mcut algorithm applied to Table 1 and Data 1-7, as dendrograms, respectively[17].

We can observe that the Mcut algorithm identifies Table 1 and Data 1-3 as dendrograms and clusters the North Kuril dialect (No. 20 in Figure 1) with the Yakumo, Oshamambe, and Soya dialects (Nos. 1, 2, and 13 in Figure 1) for Data 1 on the top left in Figure 4, with the Soya dialect (No. 13 in Figure 1) for Data 2 on the top right in Figure 4, and with the Samani, Obihiro. Kushiro, Bihoro, Soya dialects (Nos. 7-10 and 13 in Figure 1) for Data 3 on the second top left in Figure 4. Furthermore, we can also observe that the Mcut algorithm identifies Data 4-7 as dendrograms.

These results represent the purpose of sorting and changing Table 1 as explained in the Section 2.3. Therefore, I adopt Spectral Clustering as an alternative statistical method to "Large method" and "Small method" from an "identifiability" perspective, the evaluation criterion in Section 2.3. Therefore, in the subsequent analysis, I adopt the result of the Mcut algorithm applied to Table 1 (i.e., Figure 3) as a more reliable classification of Ainu dialects than the classifications through "Large method" and

intuitively. Since their enumerator and denominator (i.e., $CUT(A, A^c)$ and $W(A)$, or $CUT(B, B^c)$ and $W(B)$) share a common unit of 1 in the definition of "relation index" (Asai 1974: 61-62), their quotients (i.e., $CUT(A, A^c)/W(A)$ and $CUT(B, B^c)/W(B)$) can utilize as the measure of goodness for partition from linguistic perspectives. However, there exists no common unit on the quotients that comprises the basis on the addition and comparison of the quotients. Thus, I assume addition and comparison on the quotients. Further discussion on the issues will be promising in future.

[16] Note that, in general, the algorithm requires a vast amount of computation time as the number of objects (i.e., n) increases. For example, if researchers need to discover the 2 divisions that minimize $MCUT(A, B)$ by calculating and comparing all possible divisions, the computation time seems to increase to the fourth power as n grows. However, it has been proven that the clustering results of the Mcut algorithm coincide with the solution of a certain eigenvalue problem on the similarity matrix. These statistical properties of Spectral Clustering reduce the vast computation time and enable researchers to use Spectral Clustering as a feasible and practical method. For details regarding this issue on Spectral Clustering, the interested reader can refer to von Luxburg (2007) in the literature in English or Shinnou (2007: Ch. 7) in Japanese.

[17] The bottom right on Figure 4 also demonstrates the result of Ncut algorithm (Shi and Malik 2000), another type of spectral clustering, to Table 1 and Data 1-3, which produces the identical classification as a result. Thus, this paper does not adopt Ncut algorithm from "identifiability."

"Small method" in Figure 2[18]. Figure 5 illustrates the results of the Mcut algorithm applied to Table 1 in Figure 3 with the 12 divisions shown on the map.

Notably, Figure 5 demonstrates that the 21 Ainu dialects can be classified into two groups: Hokkaido Ainu dialects (Nos. 1-13, 20, and 21 in Figure 5) and Sakhalin Ainu dialects (Nos. 14-19 in Figure 5). Thus, the North Kuril Ainu dialect (No. 20 in Figure 5) belongs to northeastern Hokkaido Ainu dialect group (Nos. 7-13 in Figure 5). Again, note that "Major Division" that Asai (1974) established classifies the 21 Ainu dialects into three groups: Hokkaido Ainu dialects (Nos. 1-13 and 21 in Figure 1), North Kuril Ainu dialect (No. 20 in Figure 1), and Sakhalin Ainu dialects (Nos. 14-19 in Figure 1).

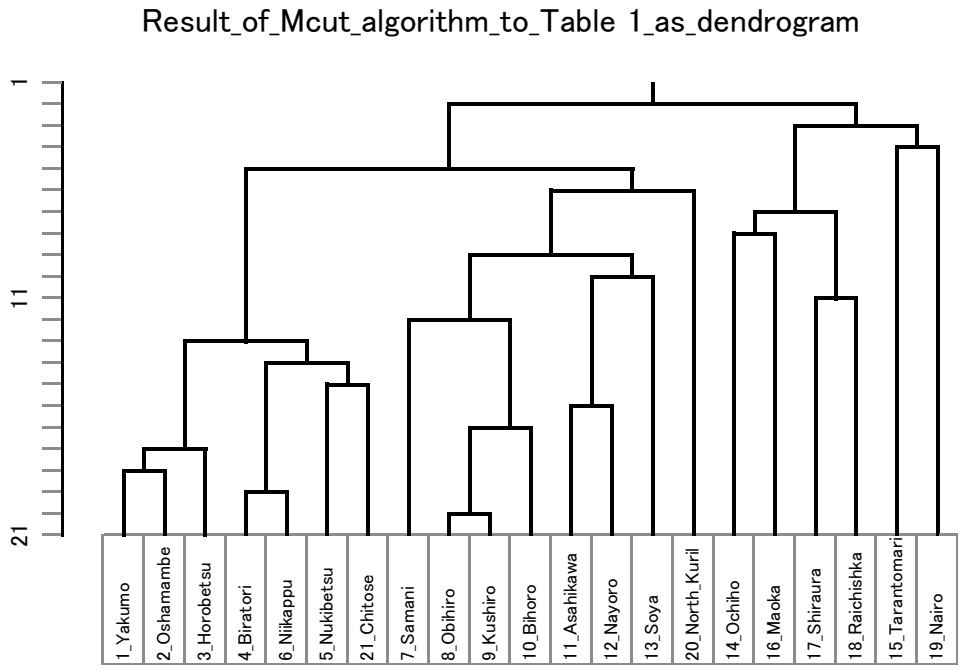

Figure 3. The result of the Mcut algorithm applied to Table 1.

[18] Asai (1974: 61) explained, "*We do not assume that the resulting tree might reflect the historical development of language under consideration. It may be useful to apply the device of cluster analysis for convenience in classifying dialects which certainly belong to one language, especially when we only have scant information other than basic words and we are obliged to analyze the data as given.*" Thus, the purpose of applying a cluster analysis (i.e., "Large method" and "Small method" in this case) to Asai (1974: 92; Table 1) is considered to be the classification of the Ainu dialects with the most feasible method but not the construction of the phylogeny of the Ainu dialects with statistical methods.

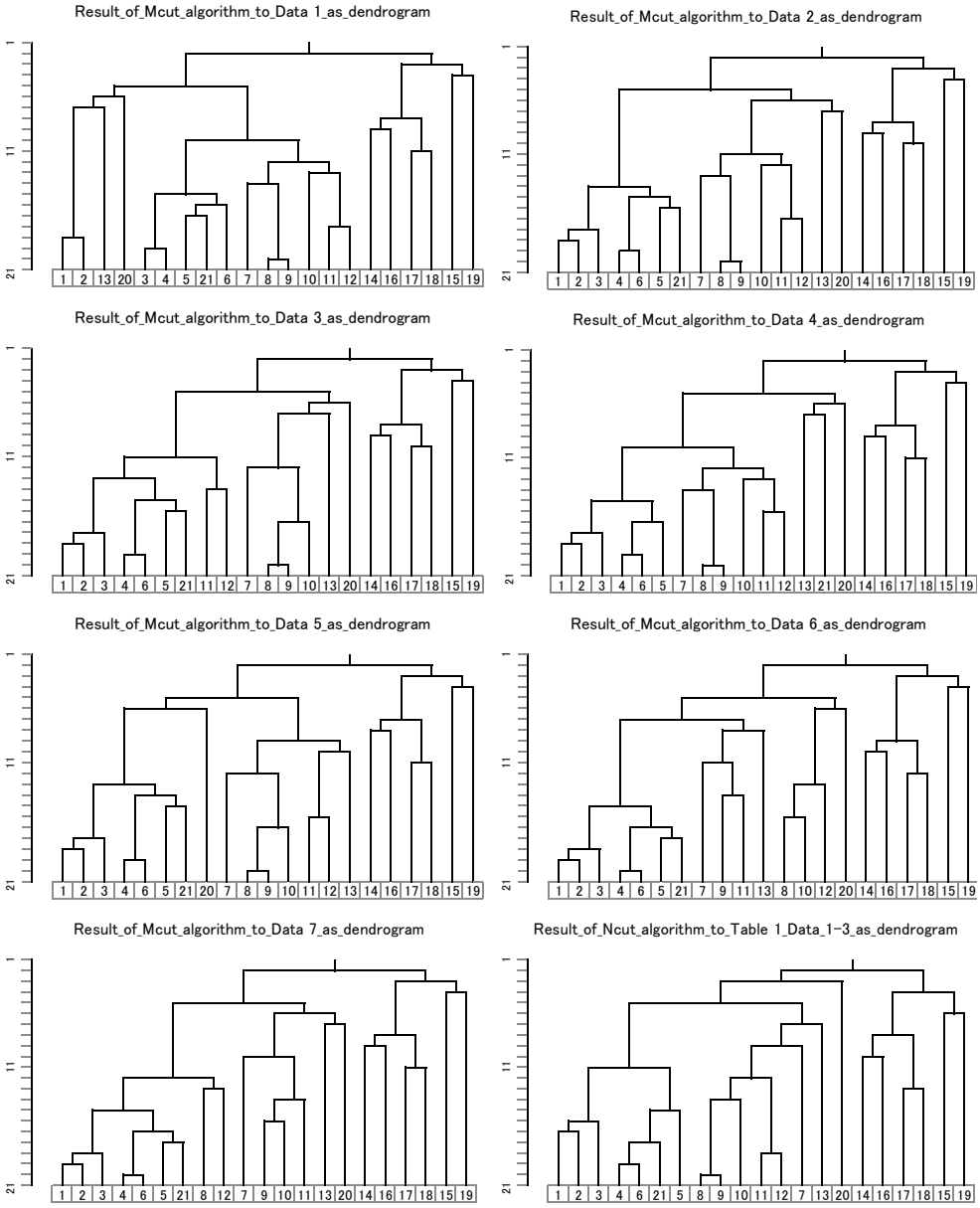

Figure 4. The results of the Mcut algorithm applied to each data.
Top left: Data 1; Top right: Data 2; Second top left: Data 3; Second top right: Data 4;
Second bottom left: Data 5; Second bottom right: Data 6; Bottom left: Data 7;
Bottom right: the result of Ncut algorithm to Table 1 and Data 1-3.
Note that the numbers in each figure correspond to the places in Figure 1.

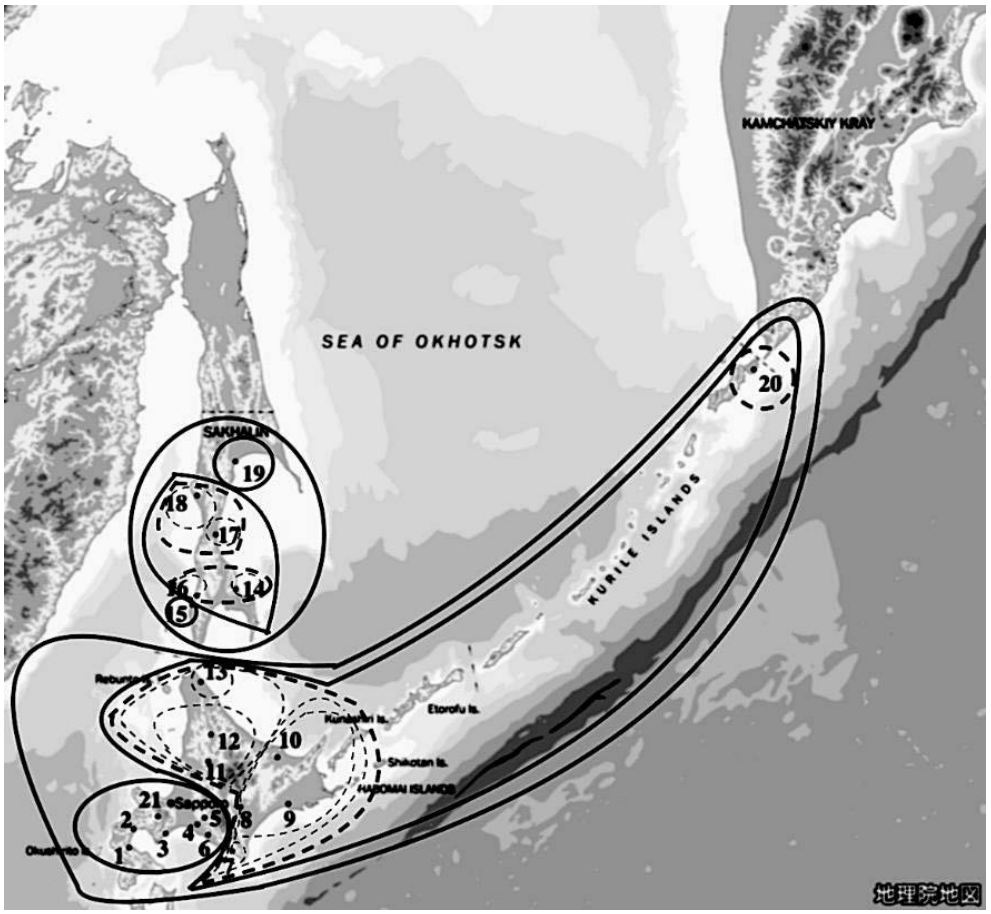

Figure 5. Map on the result in Figure 3. The number of divisions in the Mcut algorithm is 12.

Furthermore, the North Kuril dialect also belongs to southwestern Hokkaido Ainu dialect group (Nos. 1-6, and 21 in Figure 5) in the second strongest structure among the Hokkaido Ainu dialects (Nos. 1-13, 20, and 21 in Figure 5), under which the objective function of Mcut algorithm is the second minimal on the computer calculation[19,20].

4. Discussion and Conclusions

In this section, I discuss what the main results in this paper suggest for current and

[19] The objective function of Mcut algorithm can be calculated about all partitions on dialect groups.

[20] Moreover, the sub-classification of the Ainu dialects in Figure 5 exhibits a different structure from Asai's (1974) "Minor Division" in several aspects. For example, East Hokkaido (Nos. 8-12 in Figure 1) splits into two groups: the Obihiro, Kushiro, and Bihoro dialects, and the Asahikawa and Nayoro dialects, the latter of which forms a cluster with the Soya dialect. These statistical findings on East Hokkaido (Nos. 8-12 in Figure 1) are consistent with previous statistical analyses (Ono 2019b) utilizing Hattori and Chiri's (1960) lexicostatistical survey and their cognacy judgments.

future linguistic research from the viewpoints of both statistics and linguistics.

From a statistical viewpoint, the main results of this study can be summarized in three points. First, the main results in this paper shed new insight into Spectral Clustering. Although some researchers (Wieling and Nerbonne 2011) have already introduced Spectral Clustering into dialectology, those studies concentrated on the performance of Spectral Clustering compared to other clustering methods, but not on the potential of Spectral Clustering to be applied to other types of data.

The main result in this study, that Spectral Clustering is a suitable method for analyzing data that contain only information on the similarity among objects, suggests a need to reconsider previous lexicostatistical research that utilized the information on the common words among dialects or languages similarly to Asai (1974).

Moreover, the scope of this attempt could extend beyond lexicostatistics. Since the “similarity” (or “dissimilarity”) information is often only of concern in linguistic research, the application of Spectral Clustering has the potential to contribute to these fields.

Second, the success of graph-theoretic approaches (i.e., Spectral Clustering in this study) in dialectology presents at least three challenging problems for similarity data: (1) the simultaneous visualization of conflicting signals in similarity data (e.g., Network Analysis based on similarity), (2) the simultaneous visualization of the entire relationship among dialects (or languages) and words (or word forms), and (3) the extraction of the words (or the word forms) corresponding to the signals visualized in (1).

Finally, there still remain the disputed issues on the choice of “similarity” or “dissimilarity” indices in the realm of linguistics, including lexicostatistics (Asai 1974: 53-58). However, to my knowledge, those studies have not clearly focused on the relation between the index of similarity or dissimilarity and its mathematical properties (i.e., what operations are meaningful from a linguistic perspective).

As the main results of this study demonstrate, the choice of a similarity or dissimilarity index that is admissible for the linguistic environments in each dialect or language restricts the applicable operations (e.g., addition and comparison is allowed for Asai [1974: 92; Table 1] but subtraction is not applicable from a linguistic perspective). Therefore, researchers need to pay more attention to the problem of whether the similarity (or dissimilarity) index they chose violates the assumptions of the statistical analysis they choose to apply (e.g., the violation mentioned in non-metric and metric Multidimensional Scaling in this study).

These facts indicate that linguistic researchers need to distinguish between two cases: researchers can apply any statistical methods to linguistic data, which may be meaningless from a linguistic perspective, and researchers should apply statistical methods to linguistic data that are meaningful from a linguistic perspective.

The distinction will be a topic of future interdisciplinary research among the humanities and statistics.

From a linguistics viewpoint, the main results in this study can be summarized by three points. First, the "Major Division" of Ainu dialects, which classifies them into three groups—Hokkaido Ainu dialects (Nos. 1-13 and 21 in Figure 1), the North Kuril Ainu dialect (No. 20 in Figure 1), and Sakhalin Ainu dialects (Nos. 14-19 in Figure 1)—needs to be reconsidered, so far as researchers recognize the classification of Ainu dialects based on the clustering results in Asai (1974).

Rather, the 21 Ainu dialects can classify into two groups: Hokkaido Ainu dialects (Nos. 1-13, 20, and 21 in Figure 5), and Sakhalin Ainu dialects (Nos. 14-19 in Figure 5). Notably, the North Kuril Ainu dialect (No. 20 in Figure 5) belongs to the northeastern Hokkaido Ainu dialect group (Nos. 7-13 in Figure 5) in the strongest structure, under which the objective function of Mcut algorithm is minimal among the Hokkaido Ainu dialect (Nos. 1-13, 20, and 21 in Figure 5), and also belongs to the southwestern Hokkaido Ainu dialect group (Nos. 1-6, and 21 in Figure 5) in the second strongest structure, under which the objective function of Mcut algorithm is the second minimal among the Hokkaido Ainu dialect (Nos. 1-13, 20, and 21 in Figure 5).

Since the disposition of the North Kuril dialect among the Ainu language is not only of concern for the classification of Ainu dialects but also matters in the origin and dispersion of the Ainu language, further linguistic investigations will be promising[21].

Moreover, the reconsideration of the "Major Division" of Ainu dialects suggests a need for review of linguistic research based on the clustering results in Asai (1974) and

[21] Notably, Murayama (1971: 130) already noted, before Asai (1974), that "the lexical analyses illustrate that North Kuril Ainu dialect is very closely related to North Hokkaido Ainu dialects but, at the same time, shares a few common features with Sakhalin Ainu dialects. The author considers this fact and its interpretation. With the consideration of the fact that the Sea of Okhotsk is between North Kuril and Sakhalin, it might be impossible to argue that the contact between the two dialects produced common linguistic features between them. In my opinion, the interpretation explained below is reasonable: (1) Hokkaido Ainu dialects split into a south and north group; (2) a part of Ainu who speak the north dialect transferred to Sakhalin at the early stage; (3) the other part of Ainu who speak the north dialect transferred to North Kuril Island after (2) and North Kuril dialect was formed there. Thus, the author presents the possibility that both the Sakhalin Ainu dialects and the North Kuril Ainu dialect are produced from North Hokkaido Ainu dialects." Furthermore, Satō and Bugaeva (2019) have investigated unpublished 19th Kuril Ainu documents and stated that "*The lexicon of the KS [Kuril'skie Slova] variety of Kuril Ainu is mixed: some items cluster with Southern Hokkaido (Saru/Chitose) and others with Eastern Hokkaido and Sakhalin Ainu, which among other features show a pa: ca [tʃa] correspondence in a limited number of words, e.g. pa 'head', patoy 'lips', parunpe 'tongue', pas 'run' vs. ca 'head', catoy 'lips', carunpe 'tongue', cas 'run', see details on the distribution in Fukazawa (2014, 2016). KS is inconsistent with regard to the pa: ca correspondence; for instance, it has pa for 'head' (#199) but car for 'mouth' (#980), catoy for 'lips' (#160-162) and cas for 'run' (#25)*" (Satō and Bugaeva 2019: 82). As a result, the statistical findings in this paper coincide with both Murayama's insightful views and Satō and Bugaeva's (2019) invaluable contribution on North Kuril.

further philological investigations of Ainu linguists' statement about the North Kuril dialect before and after Asai (1974).

Second, the "Minor Division" of the Ainu dialects also needs to be reconsidered, so far as researchers recognize the classification of Ainu dialects based on the clustering results in Asai (1974). The sub-classification of the Ainu dialects obtained by Spectral Clustering displays a different structure than Asai's (1974) "Minor Division" in several aspects, especially for the "Northeastern" Hokkaido Ainu dialects (Nos. 8-12 in Figure 1) and the Soya dialect (No. 13 in Figure 1).

The reconsideration of the classification of Ainu dialects based on Hattori and Chiri (1960) and Asai (1974) will be of concern from the viewpoints of linguistics and statistics in future research. A reader interested in the statistical reconsideration can refer to Ono (2019a, 2019b).

Third, the results in this study indicate the need for further investigation of old documentation on the North Kuril Ainu dialects. Since Murasaki (1963) reported the extinction of the Kuril Ainu dialects, research on old documentation on (North) Kuril dialects is indispensable for Ainu linguistics. Recent linguistic and philological research (Satō and Bugaeva 2019) addresses this issue.

Finally, it is possible that "Major Division" (and "Minor Division"), which Asai (1974) may not have intended to "establish" but rather has attempted to apply clustering methods as a reference for further linguistic research, resulted in greatly influencing current Ainu linguistics beyond the purpose of Asai himself[22].

Since current Ainu linguistics has experienced great advances from Hattori and Chiri (1960) and Asai (1974), I hope that complementary studies between Ainu linguistics and statistics will be again more constructive and productive than previous researchers' contributions.

Acknowledgments

Parts of this paper on statistical methodologies in Section 2 were presented at the 63rd annual conference of The Mathematical Linguistic Society of Japan, September 21. I am very grateful to the conference organizers, the chairman of the session at my presentation, Prof. Tsunao Ogino (Nihon University), and participants of the conference for their questions and comments. Also, 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.

[22] Again, Asai (1974: 61) stated: "*We do not assume that the resulting tree might reflect the historical development of language under consideration. It may be useful to apply the device of cluster analysis for convenience in classifying dialects which certainly belong to one language, especially when we only have scant information other than basic words and we are obliged to analyze the data as given.*"

References

Asai, Tōru (1974) Classification of dialects: Cluster analysis of Ainu dialects. *Bulletin of the Institute for the Study of North Eurasian Culture,* 8, 45-136.

Black, Paul (1973) Multidimensional scaling applied to linguistic relationships. *Cahiers de l'Institut de Linguistique de Louvain,* 3, n5-6.

Ding, Chris, He Xiaofeng, Zha Hongyuan, Gu Ming and Simon Horst (2001) A min-max cut algorithm for graph partitioning and data clustering. In Cercore, Nick, Lin Tsauyoung and Wu Xindong (eds.), *Proceedings of the first IEEE International Conference on Data Mining (ICDM),* 1, 107–114. Washington: IEEE Computer Society, USA.

Dyen, Isidore, Kruskal Joseph B. and Black Paul (1992) An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society,* 82(5), 1-132.

Everitt, Brian (1979) Unresolved problems in cluster analysis. *Biometrics,* 35(1), 169-181.

Geospatial Information Authority of Japan (2019) Ministry of Land, Infrastructure, Transport and Tourism. URL: https://maps.gsi.go.jp [accessed on March 2019].

Gondran, Michel and Minoux Michel (2008) *Graphs, Dioids and Semirings: New Models and Algorithms*. New York: Springer Science+Business Media.

Hattori, Shirō and Chiri Mashiho (1960) Ainugo shohōgen no kisogoi tōkeigakuteki kenkyū [A lexicostatistical study on Ainu dialects]. *Kikan minzokugaku kenkyū [The Japanese Journal of Ethnology],* 24(4), 307-342.

Kindaichi, Kyōsuke (1932) Hokuou chimeikou: Ōu no chimei kara mita Honshū ezogo no kenkyū [A philological study of place names in the north part of Ōu region]. In Kanazawa hakase kanreki kinen shukugakai (eds.), *Tōyōgogaku no kenkyū: Kanazawa hakase kanreki kinen [A Festschrift for Kanazawa Shōzaburō in honor of his sixth birthday],* 459-552, Tokyo: Sanseidō.

Kruskal, Joseph B. (1964a) Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika,* 29(1), 1-27.

Kruskal, Joseph B. (1964b) Nonmetric multidimensional scaling: A numerical method. *Psychometrika,* 29(2), 115-129.

Murasaki, Kyōko (1963) Chishima Ainu-go zetsumetsu no hōkoku [A report on the extinction of Kuril Ainu dialects]. *Kikan minzokugaku kenkyū [The Japanese Journal of Ethnology],* 27(4), 657-661.

Murayama, Shichirō (1971) *Kita Chishima Ainu-go [Ainu Language of Northern Kuril Islands]*. Tokyo: Yoshikawa-Kōbun-Kan.

Ng, Andrew Y., Jordan Michael I. and Weiss Yair (2002) On spectral clustering: Analysis and an algorithm. In Dietterich, Thomas G., Becker Suzanna and Ghahramani Zoubin (eds.), *Advances in Neural Information Processing Systems,*

14, 849-856. Cambridge: MIT Press.
Ono, Yohei (2019a) The Ordinal Scale on Lexicostatistical Data in Ainu Dialects: Towards a New Interdisciplinary Research among the Humanities and Statistics. *Journal of the Center for Northern Humanities,* 12, 89-110.
Ono, Yohei (2019b) Observations on “Northeastern” Hokkaido Ainu Dialects: A Statistical Perspective. *Northern Language Studies,* 9, 95-122.
Ono, Yohei (2020, to appear) Some Remarks on Cognacy Judgments of Ainu Dialects: On Asai (1974). *Journal of the Center for Northern Humanities*, 13.
R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL: https://www.R-project.org/.
Satō, Tomomi and Bugaeva Anna (2019) The Study of Old Documents of Hokkaido and Kuril Ainu: Promise and Challenges. *Northern Language Studies,* 9, 67-93.
Shi, Jianbo and Malik Jitendra (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(8), 888-905.
Shinnou, Hiroyuki (2007) *R de manabu kurasuta kaiseki [Learning cluster analysis with R]*. Tokyo: Ohmsha.
Sokal, Robert R. and Michener Charles D. (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin,* 38, 1409–1438.
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.
Torii, Ryūzō (1903) *Chishima Ainu [Kuril Ainu]*. Tokyo: Yoshikawa-Kōbun-Kan.
von Luxburg, Ulrike (2007) A tutorial on spectral clustering. *Statistics and Computing,* 17(4), 395-416.
Ward Jr, Joe H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* 58(301), 236-244.
Wieling, Martijn and Nerbonne John (2011) Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech & Language,* 25(3), 700-715.

Summary

The main objective of this study is to reconsider Asai’s (1974) “Major Division” of Ainu dialects from the viewpoint of statistical methodologies. “Major Division”, which classifies Ainu dialects into the Hokkaido Ainu dialects, North Kuril Ainu dialect, and Sakhalin Ainu dialects, greatly influences current Ainu linguistics. Since there do not exist any other studies where Ainu linguists demonstrate the cognacy judgments among these Ainu dialects, Asai’s (1974) study is still a landmark in Ainu dialectology.

However, as a statistician, I was confronted with a question regarding the statistical methods used in Asai (1974). There were other clustering methods considered as able to

produce more reliable classifications than “Large method” and “Small method” in Asai (1974), which correspond to the single linkage method and complete linkage method (Sørensen 1948), respectively.

As I demonstrate in this study, Asai’s (1974: 92; Table 1) data contain only information on the similarity among the Ainu dialects but do not preserve the information on the dissimilarity among the Ainu dialects, or even the order of dissimilarity. Therefore, these statistical properties of Asai (1974: 92; Table 1) violate the assumptions of classical statistical methods, including metric and non-metric Multidimensional Scaling. Furthermore, several descriptions in Asai (1974) indicate that Asai himself noticed the particularity of his data.

This study proposes an alternative approach: Spectral Clustering, a type of graph-theoretic method that can be applied to the data structure present in Asai (1974). The main results of this study demonstrate that the “Major Division” in Ainu dialects needs to be reconsidered, so far as researchers recognize the classification of Ainu dialects based on Asai’s (1974) clustering results. Rather, the Ainu dialects can be classified into two groups: Hokkaido Ainu dialects and Sakhalin Ainu dialects. Furthermore, the North Kuril Ainu dialect belongs to the northeastern Hokkaido Ainu dialect group in the strongest structure and to southwestern Hokkaido Ainu dialect group in the second strongest structure among the Hokkaido Ainu dialects.

The statistical finding on the North Kuril Ainu dialect agrees with the philological research in both Murayama (1971) and Satō and Bugaeva (2019). These facts suggest that Asai’s (1974) “Major Division” resulted in a great influence on current Ainu linguistics, beyond Asai’s initial purpose.

(linguistics.dialectometry@gmail.com)