HOKKAIDO UNIVERSITY

Title	Statistical Remarks on Classification of Ainu Dialects : Lexicostatistics Revisited
Author(s)	Ono, Yohei
Citation	北方言語研究, 12, 185-210
Issue Date	2022-03-20
DOI	https://doi.org/10.14943/101902
Doc URL	https://hdl.handle.net/2115/84896
Type	departmental bulletin paper
File Information	13_Ono.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Northern Language Studies 12: 185-210 (Japan Association of NoLS, 2022)

Statistical Remarks on Classification of Ainu Dialects:

Lexicostatistics Revisited

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

Keywords: Ainu, ABA-type distribution, definition of numbers, lexicostatistics, missing values

1. Introduction

1.1 Motivation

There are two landmark lexicostatistical studies in Ainu dialectology focused on the classification of Ainu dialects, one by Hattori and Chiri (1960) and the other by Asai (1974b). For half a century, researchers needed to propose a classification of Ainu dialects, integrating Hattori and Chiri's and Asai's results.

Since lexicostatistical studies normally consist of cognacy judgments and statistical methodologies, the proposed classification of Ainu dialects will reflect advanced cognacy judgments[1] and current statistical methodologies. However, statisticians inherently have some limitations on their lexicostatistical research.

If statisticians were to perform cognacy judgments, their research would consist of their own cognacy judgments and statistical methodologies, resulting in no comparability with Hattori and Chiri (1960) and Asai (1974b). Further, I believe that cognacy judgment by statisticians exceeds the bounds of what they are expected to do. Therefore, in order to validate Hattori and Chiri (1960) and Asai (1974b), lexicostatistical research by statisticians should follow cognacy judgments in either of the previous studies.

As shown below, lexicostatistical studies in Ainu dialects have achieved some progress in recent years. For example, Ono (2020c) has revealed a clear disposition between Hattori and Chiri (1960) and Asai (1974b) on cognacy judgments, and Ono (2020b) has applied novel statistical methodologies to Asai's cognacy judgments, reconsidering the "major division" of Ainu dialect and illustrating its alternative.

Thus, this paper tentatively broadens my lexicostatistical contribution to Ainu dialectology by applying the statistical analysis in Ono (2020b) to Hattori and Chiri,

[1] In the following sentences, unless italicized, the English translation of the Japanese literature is by the author. Cognacy judgment is a judgment whether the two word forms (See Footnote 6) are cognate or not. The definition of cognate is as follows: "*A language or a linguistic form which is historically derived from the same source as another language/form, e.g., Spanish/Italian/French/Portuguese are 'cognate languages' (or 'cognates'); père/padre, etc. ('father') are 'cognate words' or cognates*" (Crystal 2011: 104).

comparing with Hattori and Chiri (1960) in statistical methodologies and Ono (2020b) in cognacy judgments. Notably, the discussions unfolded a critical problem—which similarity should be applied to lexicostatistical data with respect to the definition of numbers?

Although lexicostatistics is declining in linguistics, statistical analyses have experienced a great advance since the inception of this field, which could potentially contribute to linguistics by reconsidering previous research.

In advance, the problems (i.e., Problem A and Problem B) addressed in this paper are outlined in Figure 1. Note that this paper mainly focuses on Problem B in Figure 1[2,3].

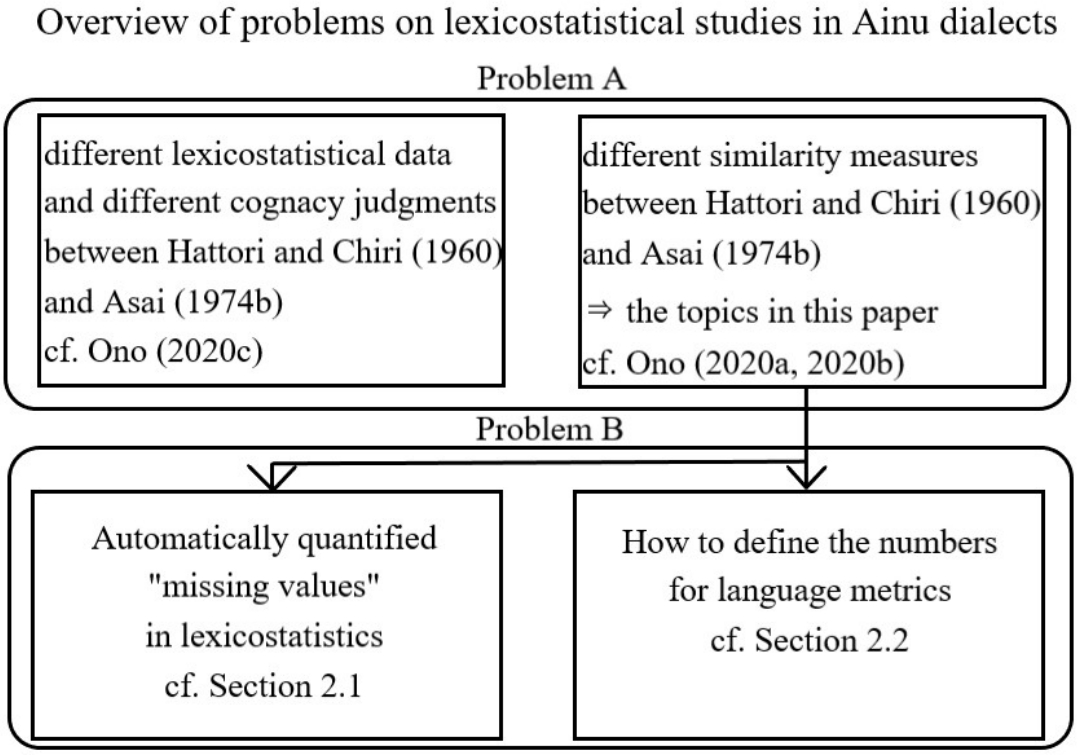

Figure 1. Overview of problems on lexicostatistical studies (i.e., Hattori and Chiri [1960] and Asai [1974b]) in Ainu dialects. Note that Ono (2020a, 2020b, 2020c) have already proposed an alternative on Problem A in Asai (1974b), and the main objective of this paper is to clarify the Problem B.

[2] One of reviewers pointed out that this paper gave the impression of being a statistical study on the Ainu language rather than a linguistic study. In my previous research, I have deliberately avoided justifying the results of statistical analysis from linguistics. Otherwise, the discussion will be circular. Thus, the objective of my previous studies is to establish a new discipline, "mathematical humanities", which investigates the mathematical properties of humanities data, selects the corresponding statistical methods, and aims to present more valid results, which could potentially contribute to further developments in the humanities. Interested reader will refer to Ono (2022b, to appear) in Japanese.

[3] In the following sentence, Hattori and Chiri (1960) and Asai (1974b) are frequently cited. For the sake of space limitation, Hattori and Chiri (1960) are abbreviated as Hattori and Chiri, and Hattori and Chiri (1960: pages) as Hattori and Chiri (pages) or (pages) if it is obvious that Hattori and Chiri (1960) are mentioned from the context. The same abbreviation applies to Asai (1974b).

1.2 Previous lexicostatistical studies and their problems

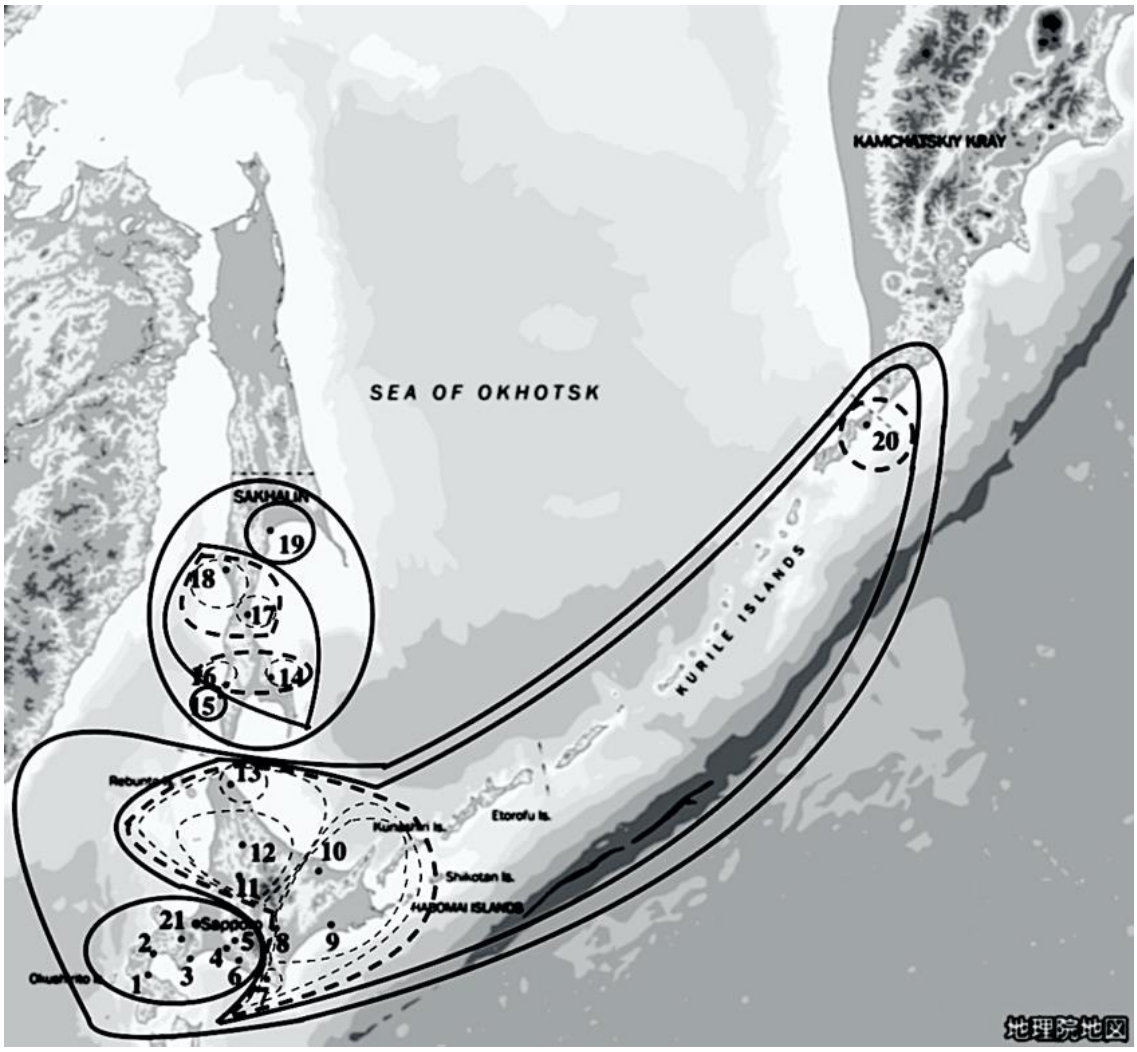

Figure 2. Map of a section of the region where the Ainu language is or was spoken and its classification cited from Ono (2020b: 248; Figure 5). The dotted and solid lines correspond to the hierarchy of dialect groupings in Ono (2020b: 246; Figure 3) with 12 divisions. Note that 1: Yakumo, 2: Oshamambe, 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.

Hattori and Chiri's study consisted of a lexicostatistical survey of 19 Ainu dialects (Nos. 1–19 in Figure 2), with reference to Swadesh's (1955) basic word list. As Ainu dialects were starting to disappear, his research contributed considerably to Ainu linguistics.

Furthermore, Asai revised some of the lexicostatistical data in Hattori and Chiri's study of the Obihiro, Kushiro, and Asahikawa dialects (Nos. 8, 9, and 11 in Figure 2, respectively) based on his own fieldwork (66), collected the Chitose dialect (No. 21 in

Figure 2) from informants (64–66), and gathered the North Kuril dialect (No. 20 in Figure 2) by referring to Torii (1903), Murayama (1971), and Pinart's vocabulary manuscripts (Asai 1974b: Appendix). Moreover, Asai referenced Chiri (1953, 1954) and Hattori (1964) as needed.

Both Hattori and Chiri (1960) and Asai (1974b) applied statistical methods to their data and presented a classification of Ainu dialects. For example, Hattori and Chiri (338) calculated similarity among the 19 Ainu dialects[4] and summarized the relationship qualitatively in six points[5].

Further, Asai (1974b) performed cognacy judgments on 110 words among the 21 Ainu dialects (Nos. 1–21 in Figure 2)[6], calculated a similarity matrix (92; Table 1) based on his "relation index" (61–62), and applied a cluster analysis to the matrix. This classification had a great influence on contemporary Ainu linguistics. A previous study (Ono 2020b) scrutinized the properties of Asai's data from linguistic and mathematical perspectives and applied spectral clustering (a graph-theoretic approach applicable to algebraic properties) to Asai's similarity matrix as illustrated in Figure 2.

Since no other documents have recorded Ainu linguists' cognacy judgments in Ainu dialects with sufficient quality and quantity, researchers' comparison, examination, and integration of these two studies from the viewpoint of Ainu linguistics have been critical in the past 50 years. However, Ainu linguists have faced two obstacles in examining, comparing, and integrating Hattori and Chiri's and Asai's works, that is, Problem A in Figure 1.

---

[4] See Table 3 in Section 2.

[5] Note that "(1) there were significant differences between the Hokkaido Ainu dialects and Sakhalin Ainu dialects; (2) the Soya dialect was the most dissimilar among the Hokkaido Ainu dialects and the most similar to the Sakhalin Ainu dialects; (3) the similarity between some of the Hokkaido Ainu dialects (i.e., Biratori, Niikappu, Nayoro, and Asahikawa) and Sakhalin Ainu dialects was slightly greater than that between the other Hokkaido Ainu dialects and Sakhalin Ainu dialects, which is of concern for future research; (4) Hokkaido Ainu dialects formed different clusters (e.g., Yakumo and Oshamambe; Biratori, Nukibetsu, and Niikappu; Obihiro, Kushiro and Bihoro). However, when we drew a sample from each cluster, they were more dissimilar to each other (e.g., Oshamambe, Biratori, Nayoro, and Bihoro); (5) the Samani dialect was significantly dissimilar to the Biratori, Nukibetsu, and Niikappu dialects, corresponding to the 'customary difference' between southern and northern Hidaka. Prof. Chiri told me that, in Ainu folklore, there was a great war between them. Note that the Samani dialect belonged to the former and the Biratori, Nukibetsu, and Niikappu dialects belonged to the latter; (6) the Samani dialect was more similar to the Obihiro and Kushiro dialects, which is of concern for future research" (Hattori and Chiri 1960: 399–340).

[6] In this paper, I use "the word" as the entry word in the basic word list (e.g., 'here' in Table 2) and "the word form" as the specific word representing "the word" in a certain dialect (e.g., *téta* in the Yakumo dialect), according to the notation in Asai (1974b). However, in general, "the word" corresponds to the lexeme or the citation form in linguistics (e.g., WALK in English) and "the word form" corresponds to the particular form of "the word" (e.g., walk, walks, walking, and walked in English). Interested readers need to pay attention to the different meanings of "the word" and "the word form" when referring to Asai (1974b).

Table 1. Non-cognate judgments by Asai and cognate judgments by Hattori and Chiri, the former of which is in Ono (2020c: 53). The numbers in parentheses after each word form correspond to the places in Figure 2.

basic_word_list	specified non-cognate judgments in Asai (1974)	Hattori and Chiri's (1960) cognacy judgments
19.fish	cep(3-13, 15, 19-21)/ceh(14, 16-18)	cognate
36.feather	rap(1-13, 15, 20, 21)/rah(14, 16-18)	cognate
43.tooth	imak(7-10, 13, 15, 19, 20)/imah(14, 16-18)	cognate
48.hand	tek(1-13, 15, 19-21)/teh(14, 16, 18)	cognate
50.neck	rekut(1-13, 15, 20, 21)/rekuh(14, 16-18)	cognate
66.come	ek(1-13, 15, 19-21)/eh(14, 16-18)	cognate
93.hot	sesek(1-13, 15, 19-21)/seseh(14, 16-18)	cognate
129.wing	tekkup(3, 4, 6, 8-12, 15, 19, 20)/tehkuh(16, 18)	cognate
146.wife	mat(1-13, 15, 19-21)/mah(14, 16-18)	cognate
147.salt	sippo(1-13, 15, 19, 21)/sispo(14, 16-18)	cognate
148.ice	rup(12, 13, 15)/ruh(14, 16-18)	cognate
198.alive	siknu(3-9, 11, 12, 15, 20, 21)/sisnu(14, 16-19)	cognate

First, as Ono (2020c) demonstrated, Hattori and Chiri (1960) and Asai (1974b) analyzed different lexicostatistical data and adopted different cognacy judgments, although the latter were not specified prior to the recent study (Ono 2020c) in which they were partially revealed.

Table 1 illustrates the cognacy judgments in the two studies. Hattori and Chiri's cognacy judgments reflect the following description in Hattori (1967: 208–209): "CVw, CVy, CVm, CVn, and CVs agree in both the Hokkaido and Sakhalin Ainu dialects. Furthermore, CVp, CVt, and CVk in the Hokkaido Ainu dialects correspond to CVh in the Sakhalin Ainu dialects and CVr in the Hokkaido Ainu dialects to CVrV in the Sakhalin Ainu dialects. Note that Cip, Cit, and Cik in the Hokkaido Ainu dialects correspond to Cis in the Sakhalin Ainu dialects." However, Table 1 clearly demonstrates that Asai adopted Hattori and Chiri's judgment not as cognate but as non-cognate[7,8].

Second, Hattori and Chiri (1960) and Asai (1974b) measured similarity among Ainu dialects using different methods. Table 2 shows the word form on 'here' in Hattori and Chiri (314) and the corresponding cognacy judgments between the Yakumo dialect record (i.e., *téta*) and the others by Hattori and Chiri.

Hattori and Chiri (307) introduced seven symbols for the cognacy judgments in Ainu dialects, which indicated some unavoidable uncertainty in the linguistic environment

[7] Moreover, Ono (2020c) demonstrated that Asai judged on 'back' *seturu* (14, 16-20) and *seturi* (15) and on 'ye' *ecookaj utara* (16, 19) and *ecookaj utari* (15) as non-cognate.

[8] Asai (1974b) did not provide any specific cognacy judgment or explain cognacy judgments from Ainu linguistics. However, Asai (1974)'s classification of Ainu dialects still has great influence on current Ainu linguistics. Thus, it is of concern to evaluate Asai's cognacy judgments, and to verify Asai's results from current Ainu linguistics. I am currently working on this issue in collaboration with an Ainu linguist, and we plan to present our research as articles in this year.

Table 2. Word forms of 'here' in Hattori and Chiri (324) and the corresponding cognacy judgments between the record of Yakumo dialect (i.e., *téta*) and the others.

	Word forms on 'here' in Hattori and Chiri (1960)	Cognacy Judgments based on Yakumo dialect record in Hattori and Chiri (1960)
Dialect	Word Forms	Cognacy Judgments
1_Yakumo	téta	+
2_Oshamambe	téta	+
3_Horobetsu	téta	+
4_Biratori	téta	+
5_Nukibetsu	téta	+
6_Niikappu	té'or	( )
7_Samani	ta'anta	−
8_Obihiro	ta'ánta	−
9_Kushiro	tanta	−
10_Bihoro	temanta	◯
11_Asahikawa	téta	+
12_Nayoro	téta, tánta	±
13_Soya	téta	+
14_Ochiho	teeta	+
15_Tarantomari	teeta	+
16_Maoka	teeta	+
17_Shiraura	teeta	+
18_Raichishka	teeta	+
19_Nairo	teyta	+

surrounding the Ainu language in the 1960s, as follows: +: "*cognate residues*"; −: "*non-cognates*"; ±: "*cognates and non-cognates (when one or both the dialects have two forms, and the imperfectness of the record does not allow us to decide which is more basic)*"; ○: "*questionable etymology or choice*"; ?: "*doubtful record*"; · : "*no answer given*"; (): "*lacuna of record.*"

Further, as discussed in Section 2, Hattori and Chiri (337) calculated similarity between pairs of dialects by dividing the numerator (the number of + and ±, the latter of which is weighted by the coefficient [i.e., 0.5]) by the denominator (the number of +, ±, and − in pairs of dialects), whereas Asai (1974b) calculated similarity between pairs of dialects by counting the number of + and ± in the definition of his "relation" index (Asai 1974b: 61–62).

For example, the cognacy judgments between the Yakumo and Oshamambe dialects (Nos. 1–2 in Figure 2) contain 183 +, five ±, five −, one ?, and two •. Thus, Hattori and Chiri quantified the numerator as 1*183+0.5*5=185.5, the denominator as 1*183+1*5+1*5=193, and the similarity between Yakumo and Oshamambe as 185.5/193 ≒0.961 rounded to the fourth decimal, while Asai's method described the similarity between Yakumo and Oshamambe as 1*183+1*5=188, tentatively assuming the same cognacy judgments as in Hattori and Chiri. As illustrated in this paper, the two measures lead to different classifications of Ainu dialects.

Consequently, researchers must address the two problems (i.e., Problem A in Figure 1) of Hattori and Chiri (1960) and Asai (1974b) "separately," to enable them to examine, compare, and integrate their work. As Ono (2020c) focused on the former problem in Problem A: different lexicostatistical data and different cognacy judgments between Hattori and Chiri (1960) and Asai (1974b), the main objective of this paper is to clarify the latter problem in Problem A: the two different calculation methods of Hattori and Chiri (1960) and Asai (1974b), and to verify the two different classifications from mathematical and statistical viewpoints.

Since Ono (2020c) proved that the descriptions in Asai were insufficient for specifying all cognacy judgments adopted by Asai, this paper focuses on the cognacy judgments and similarity matrix in Hattori and Chiri (338).

1.3 Summary and organization of this paper

The main results of the current study are summarized in five points. First, I report an error concerning cognacy judgment by Hattori and Chiri (1960) and propose a correction. Based on my correction, the revised figures of a previous study (Ono 2015) are also presented in the Appendix. Furthermore, I describe miscalculations in the similarity matrix by Hattori and Chiri (338) and demonstrate the corrected similarity matrix. Thus, I illustrate the recalculated similarity matrices in both Hattori and Chiri's and Asai's approaches by utilizing the corrected cognacy judgments of Hattori and Chiri. Consequently, the miscalculation causes different classifications.

Second, I examine various similarity matrices generated by the two methods from mathematical and statistical viewpoints and propose the matrix to be analyzed based on substantive knowledge in both linguistics and statistics. Notably, the analyses demonstrate two critical problems (i.e., Problem B in Figure 1) in previous lexicostatistical studies: the automatic quantification on "missing values" and the definition of numbers for language metrics[9].

Third, I apply spectral clustering, a graph-theoretic approach adopted in previous studies (Ono 2020a, 2020b), to the matrices proposed in this study. Notably, the classification of Ainu dialects obtained statistically is generally consistent with that obtained by applying the same method to Asai's data, as shown in Figures 3–4. However, the results confirm the significance of the two problems in lexicostatistical studies. Several differences between the two classifications are also reported.

Fourth, I investigate the other strongest classifications in the Hokkaido Ainu dialects (Nos. 1–13 in Figure 2), under which the objective function of spectral clustering is the second and subsequent minimal. The result leads to an ABA-type distribution and its variant, which a previous study (Ono 2020a) also reported for Asai's data using the same approach.

Finally, the main results suggest the need to reconsider the previous classification of Ainu dialects based on the statistical analyses in both Hattori and Chiri (1960) and Asai (1974b). Thus, the analyses also suggest that previous lexicostatistical data contain potentially two critical problems (i.e., Problem B in Figure 1) in the form of similarity and need to be reconsidered from the viewpoints of Asai's method.

The remainder of this paper is as follows. Section 2 shows the corrected similarity matrix in Hattori and Chiri (1960) and focuses on Problem B in Figure 1. Section 3

[9] In the following text, two problems correspond to Problem B in Figure 1.

introduces spectral clustering, a graph-theoretic approach appropriate for verifying the Problem B in previous section. Section 4 shows the classifications obtained by spectral clustering and confirms that the Problem B resulted in different results. Section 5 discusses the significance of this study from both linguistics and statistics viewpoints.

2. Materials and their problems

Table 3 demonstrates the original similarity matrix among the 19 Ainu dialects in Hattori and Chiri (338). Hattori and Chiri (312) calculated the similarity between pairs of dialects as $S_{ij}=(\#+_{ij}+0.5*\#\pm_{ij})/(\#+_{ij}+\#\pm_{ij}+\#-_{ij})$, while Asai (1974: 61–62) used $S_{ij}=(\#+_{ij}+\#\pm_{ij})$. The symbols following # indicate the number of the corresponding symbol between the pair of dialects *i* and *j*, $\{i,j = 1,2,\ldots,19\}$. I recalculated the corrected similarity matrix with the approach in Hattori and Chiri (338), as shown in Table 4[10].

Table 3. Similarity data among the 19 Ainu dialects by Hattori and Chiri (338). 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
1_Yakumo	100	96.1	91.8	90.3	88.8	89.1	88.7	87.5	88.2	87.7	89.5	86.7	84.8	73.3	70.3	70.8	71.9	70.5	72.0
2_Oshamambe	96.1	100	90.7	90.2	90.5	90.0	89.8	87.0	87.9	86.7	88.6	85.2	83.2	74.5	68.9	70.0	71.3	70.0	71.3
3_Horobetsu	91.8	90.7	100	93.1	90.6	92.0	87.9	88.5	90.6	87.9	90.3	88.2	85.1	74.2	73.2	73.6	74.6	73.8	75.4
4_Biratori	90.3	90.2	93.1	100	95.6	96.1	85.8	86.7	88.0	86.9	88.7	87.2	83.5	75.3	73.4	74.1	74.3	73.8	75.7
5_Nukibetsu	88.8	90.5	90.6	95.6	100	95.8	85.0	84.6	85.7	83.8	87.7	85.9	82.9	74.9	70.5	71.7	72.5	70.9	72.7
6_Niikappu	89.1	90.0	92.0	96.1	95.8	100	87.1	86.2	88.0	85.8	88.4	85.2	82.2	75.3	73.1	73.5	74.1	73.3	75.4
7_Samani	88.7	89.8	87.9	85.8	85.0	87.1	100	90.7	92.1	89.2	89.4	85.2	84.4	72.5	69.7	70.8	72.1	70.3	71.3
8_Obihiro	87.5	87.0	88.5	86.7	84.6	86.2	90.7	100	94.7	93.1	90.0	88.7	84.2	70.5	69.4	70.2	69.6	68.6	70.1
9_Kushiro	88.2	87.9	90.6	88.0	85.7	88.0	92.1	94.7	100	94.0	91.4	88.7	86.8	72.0	71.7	71.7	71.9	70.6	72.1
10_Bihoro	87.7	86.7	87.9	86.9	83.8	85.8	89.2	93.1	94.0	100	89.2	88.4	86.6	71.2	71.1	70.6	71.1	70.1	70.5
11_Asahikawa	89.5	88.6	90.3	88.7	87.7	88.4	89.4	90.0	91.4	89.2	100	90.8	85.4	73.4	73.6	73.9	74.2	73.4	75.1
12_Nayoro	86.7	85.2	88.2	87.2	85.9	85.2	85.2	88.7	88.7	88.4	90.8	100	87.6	73.9	73.0	72.8	73.0	72.3	73.2
13_Soya	84.8	83.2	85.1	83.5	82.9	82.2	84.4	84.2	86.8	86.6	85.4	87.6	100	79.7	78.8	79.4	78.3	77.5	76.8
14_Ochiho	73.3	74.5	74.2	75.3	74.9	75.3	72.5	70.5	72.0	71.2	73.4	73.9	79.7	100	88.8	91.1	91.7	89.6	92.4
15_Tarantomari	70.3	68.9	73.2	73.4	70.5	73.1	69.7	69.4	71.7	71.1	73.6	73.0	78.8	88.8	100	92.6	88.5	89.0	89.8
16_Maoka	70.8	70.0	73.6	74.1	71.7	73.5	70.8	70.2	71.7	70.6	73.9	72.8	79.4	91.1	92.6	100	91.6	90.3	92.3
17_Shiraura	71.9	71.3	74.6	74.3	72.5	74.1	72.1	69.6	71.9	71.1	74.2	73.0	78.3	91.7	88.5	91.6	100	92.1	93.3
18_Raichishka	70.5	70.0	73.8	73.8	70.9	73.3	70.3	68.6	70.6	70.1	73.4	72.3	77.5	89.6	89.0	90.3	92.1	100	90.5
19_Nairo	72.0	71.3	75.4	75.7	72.7	75.4	71.3	70.1	72.1	70.5	75.1	73.2	76.8	92.4	89.8	92.3	93.3	90.5	100

The gray cells are miscalculations by Hattori and Chiri. The data matrix is multiplied by 100, and then rounded to the second decimal in Table 4.

[10] The similarity matrix cannot be corrected by considering an original error in Hattori and Chiri: for the word 'how,' Hattori and Chiri (324) recorded *nekon(a)* in the Bihoro dialect (No. 10 in Figure 2) and *temana* in the Shiraura dialect (No. 17 in Figure 2) and described the cognacy judgment between the two dialects as +: "*cognate residues.*" As Hattori and Chiri (324) recorded *nékona* in the Yakumo dialect (No. 1 in Figure 2), *nekona* in the Kushiro dialect (No. 9 in Figure 2), and *nékon(a)* in the Soya dialect (No. 13 in Figure 2), whose cognacy judgment to *temana* in the Shiraura dialect is −: "*non-cognates*" systematically, the cognacy judgment on 'how' between the Yakumo and Shiraura dialects should be corrected to −: "*non-cognates.*" The revised figures of a previous study (Ono 2015) and the revised explanations in the corresponding figures are demonstrated as Figure 1-1, Figure 1-2, Figure 2-1, and Figure 2-2 in the Appendix.

Table 4. Similarity data among the 19 Ainu dialects recalculated using Hattori and Chiri's method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1_Yakumo	100	96.1	91.8	90.3	88.8	89.1	88.7	87.5	88.2	87.7	89.5	86.7	84.8	73.3	70.3	70.8	71.9	70.8	72.0
2_Oshamambe	96.1	100	91.7	90.2	90.5	90.0	89.8	87.0	87.9	86.7	88.6	85.2	83.2	74.5	68.9	70.0	71.3	70.0	71.3
3_Horobetsu	91.8	91.7	100	93.4	90.6	92.0	87.9	88.5	90.6	87.9	90.3	88.2	83.8	74.2	73.2	73.6	74.6	73.8	75.4
4_Biratori	90.3	90.2	93.4	100	95.6	96.1	85.8	86.7	88.0	86.4	88.7	87.2	83.5	75.3	73.4	74.1	74.3	73.8	75.7
5_Nukibetsu	88.8	90.5	90.6	95.6	100	95.8	85.0	84.6	85.7	83.8	87.7	85.9	82.9	74.9	70.5	71.7	72.5	70.9	72.7
6_Niikappu	89.1	90.0	92.0	96.1	95.8	100	87.4	86.2	88.0	86.0	88.3	85.2	82.2	75.3	73.1	73.5	74.1	73.3	75.4
7_Samani	88.7	89.8	87.9	85.8	85.0	87.4	100	90.7	92.1	89.2	89.4	85.2	84.4	72.5	69.7	71.3	72.1	70.3	71.2
8_Obihiro	87.5	87.0	88.5	86.7	84.6	86.2	90.7	100	94.7	93.1	90.0	88.7	84.1	70.5	69.4	70.2	69.6	68.6	70.1
9_Kushiro	88.2	87.9	90.6	88.0	85.7	88.0	92.1	94.7	100	94.0	91.4	88.7	86.3	71.9	71.7	71.7	71.9	70.6	72.1
10_Bihoro	87.7	86.7	87.9	86.4	83.8	86.0	89.2	93.1	94.0	100	89.2	88.4	86.1	71.2	71.1	70.6	71.1	70.1	70.5
11_Asahikawa	89.5	88.6	90.3	88.7	87.7	88.3	89.4	90.0	91.4	89.2	100	91.3	85.4	73.4	73.6	73.9	74.2	73.4	75.1
12_Nayoro	86.7	85.2	88.2	87.2	85.9	85.2	85.2	88.7	88.7	88.4	91.3	100	87.6	73.9	73.0	72.8	73.0	72.3	73.2
13_Soya	84.8	83.2	83.8	83.5	82.9	82.2	84.4	84.1	86.3	86.1	85.4	87.6	100	79.7	78.8	79.4	78.3	77.5	76.8
14_Ochiho	73.3	74.5	74.2	75.3	74.9	75.3	72.5	70.5	71.9	71.2	73.4	73.9	79.7	100	88.8	91.1	91.7	89.6	92.4
15_Tarantomari	70.3	68.9	73.2	73.4	70.5	73.1	69.7	69.4	71.7	71.1	73.6	73.0	78.8	88.8	100	92.6	88.5	89.0	89.8
16_Maoka	70.8	70.0	73.6	74.1	71.7	73.5	71.3	70.2	71.7	70.6	73.9	72.8	79.4	91.1	92.6	100	91.6	90.3	92.3
17_Shiraura	71.9	71.3	74.6	74.3	72.5	74.1	72.1	69.6	71.9	71.1	74.2	73.0	78.3	91.7	88.5	91.6	100	92.1	93.3
18_Raichishka	70.8	70.0	73.8	73.8	70.9	73.3	70.3	68.6	70.6	70.1	73.4	72.3	77.5	89.6	89.0	90.3	92.1	100	90.5
19_Nairo	72.0	71.3	75.4	75.7	72.7	75.4	71.2	70.1	72.1	70.5	75.1	73.2	76.8	92.4	89.8	92.3	93.3	90.5	100

Table 5. Correct denominators in similarity data on the 196 words among the 19 Ainu dialects recalculated using Hattori and Chiri's method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1_Yakumo	196	193	196	196	192	193	194	196	191	195	191	195	191	189	192	192	192	192	193
2_Oshamambe	193	196	193	193	190	190	192	193	190	192	188	193	187	186	190	190	190	190	190
3_Horobetsu	196	193	196	196	192	193	194	196	191	195	191	195	191	188	192	191	191	191	191
4_Biratori	196	193	196	196	192	193	194	196	191	195	191	195	191	188	192	191	191	191	191
5_Nukibetsu	192	190	192	192	196	189	190	192	189	191	187	192	187	185	188	187	187	187	187
6_Niikappu	193	190	193	193	189	196	190	192	187	193	188	192	188	184	188	187	187	187	187
7_Samani	194	192	194	194	190	190	196	193	190	194	189	193	189	187	190	190	190	190	191
8_Obihiro	196	193	196	196	192	192	193	196	190	195	190	194	189	188	191	191	191	191	192
9_Kushiro	191	190	191	191	189	187	190	190	196	191	186	191	186	185	187	187	187	187	188
10_Bihoro	195	192	195	195	191	193	194	195	191	196	190	194	190	189	192	192	192	192	193
11_Asahikawa	191	188	191	191	187	188	189	190	186	190	196	190	189	188	191	190	190	190	191
12_Nayoro	195	193	195	195	192	192	193	194	191	194	190	196	190	188	191	191	191	191	192
13_Soya	191	187	191	191	187	188	189	189	186	190	189	190	196	187	189	189	189	189	190
14_Ochiho	189	186	188	188	185	184	187	188	185	189	188	188	187	196	192	192	192	192	191
15_Tarantomari	192	190	192	192	188	188	190	191	187	192	191	191	189	192	196	196	196	196	196
16_Maoka	192	190	191	191	187	187	190	191	187	192	190	191	189	192	196	196	196	196	196
17_Shiraura	192	190	191	191	187	187	190	191	187	192	190	191	189	192	196	196	196	196	195
18_Raichishka	192	190	191	191	187	187	190	191	187	192	190	191	189	192	196	196	196	196	195
19_Nairo	193	190	191	191	187	187	191	192	188	193	191	192	190	191	196	196	195	195	196

Section 4 shows different classifications of Ainu dialects due to the miscalculation. Furthermore, the corresponding numerators ($\#+_{ij}+0.5*\#\pm_{ij}$) are demonstrated in Table 5, the corresponding denominators ($\#+_{ij}+\#\pm_{ij}+\#-_{ij}$) in Table 6, and the similarity by Asai's method, $\#+_{ij}+\#\pm_{ij}$, in Table 7. The following subsections demonstrate the abovementioned two problems in Table 4 from the viewpoints of linguistics and statistics[11].

[11] Hattori (1964: 21; Introduction) stated about Table 3 that "the values in the table correspond to the

Table 6. Correct numerators on similarity data on the 196 words among the 19 Ainu dialects recalculated using Hattori and Chiri's method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1_Yakumo	196.0	185.5	180.0	177.0	170.5	172.0	172.0	171.5	168.5	171.0	171.0	169.0	162.0	138.5	135.0	136.0	138.0	136.0	139.0
2_Oshamambe	185.5	196.0	177.0	174.0	172.0	171.0	172.5	168.0	167.0	166.5	166.5	164.5	155.5	138.5	131.0	133.0	135.5	133.0	135.5
3_Horobetsu	180.0	177.0	196.0	183.0	174.0	177.5	170.5	173.5	173.0	171.5	172.5	172.0	160.0	139.5	140.5	140.5	142.5	141.0	144.0
4_Biratori	177.0	174.0	183.0	196.0	183.5	185.5	166.5	170.0	168.0	168.5	169.5	170.0	159.5	141.5	141.0	141.5	142.0	141.0	144.5
5_Nukibetsu	170.5	172.0	174.0	183.5	196.0	181.0	161.5	162.5	162.0	160.0	164.0	165.0	155.0	138.5	132.5	134.0	135.5	132.5	136.0
6_Niikappu	172.0	171.0	177.5	185.5	181.0	196.0	166.0	165.5	164.5	166.0	166.0	163.5	154.5	138.5	137.5	137.5	138.5	137.0	141.0
7_Samani	172.0	172.5	170.5	166.5	161.5	166.0	196.0	175.0	175.0	173.0	169.0	164.5	159.5	135.5	132.5	135.5	137.0	133.5	136.0
8_Obihiro	171.5	168.0	173.5	170.0	162.5	165.5	175.0	196.0	180.0	181.5	171.0	172.0	159.0	132.5	132.5	134.0	133.0	131.0	134.5
9_Kushiro	168.5	167.0	173.0	168.0	162.0	164.5	175.0	180.0	196.0	179.5	170.0	169.5	160.5	133.0	134.0	134.0	134.5	132.0	135.5
10_Bihoro	171.0	166.5	171.5	168.5	160.0	166.0	173.0	181.5	179.5	196.0	169.5	171.5	163.5	134.5	136.5	135.5	136.5	134.5	136.0
11_Asahikawa	171.0	166.5	172.5	169.5	164.0	166.0	169.0	171.0	170.0	169.5	196.0	173.5	161.5	138.0	140.5	140.5	141.0	139.5	143.5
12_Nayoro	169.0	164.5	172.0	170.0	165.0	163.5	164.5	172.0	169.5	171.5	173.5	196.0	166.5	139.0	139.5	139.0	139.5	138.0	140.5
13_Soya	162.0	155.5	160.0	159.5	155.0	154.5	159.5	159.0	160.5	163.5	161.5	166.5	196.0	149.0	149.0	150.0	148.0	146.5	146.0
14_Ochiho	138.5	138.5	139.5	141.5	138.5	138.5	135.5	132.5	133.0	134.5	138.0	139.0	149.0	196.0	170.5	175.0	176.0	172.0	176.5
15_Tarantomari	135.0	131.0	140.5	141.0	132.5	137.5	132.5	132.5	134.0	136.5	140.5	139.5	149.0	170.5	196.0	181.5	173.5	174.5	176.0
16_Maoka	136.0	133.0	140.5	141.5	134.0	137.5	135.5	134.0	134.0	135.5	140.5	139.0	150.0	175.0	181.5	196.0	179.5	177.0	181.0
17_Shiraura	138.0	135.5	142.5	142.0	135.5	138.5	137.0	133.0	134.5	136.5	141.0	139.5	148.0	176.0	173.5	179.5	196.0	180.5	182.0
18_Raichishka	136.0	133.0	141.0	141.0	132.5	137.0	133.5	131.0	132.0	134.5	139.5	138.0	146.5	172.0	174.5	177.0	180.5	196.0	176.5
19_Nairo	139.0	135.5	144.0	144.5	136.0	141.0	136.0	134.5	135.5	136.0	143.5	140.5	146.0	176.5	176.0	181.0	182.0	176.5	196.0

Table 7. Numerators in similarity data on the 196 words among the 19 Ainu dialects recalculated using Asai's method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1_Yakumo	196	188	185	182	173	174	174	176	171	175	175	176	167	141	140	141	141	140	144
2_Oshamambe	188	196	182	177	172	171	173	170	168	168	169	170	159	139	133	136	136	135	137
3_Horobetsu	185	182	196	187	179	182	174	177	177	177	178	178	166	143	146	146	147	146	149
4_Biratori	182	177	187	196	187	189	169	174	171	173	175	176	166	144	146	146	145	145	149
5_Nukibetsu	173	172	179	187	196	181	162	165	163	162	167	171	159	139	135	137	136	135	138
6_Niikappu	174	171	182	189	181	196	166	168	165	168	169	169	158	139	140	140	139	139	143
7_Samani	174	173	174	169	162	166	196	178	175	175	172	171	162	136	135	138	138	135	138
8_Obihiro	176	170	177	174	165	168	178	196	182	185	176	178	164	135	138	139	136	135	139
9_Kushiro	171	168	177	171	163	165	175	182	196	181	174	176	163	134	137	137	136	134	138
10_Bihoro	175	168	177	173	162	168	175	185	181	196	174	179	168	137	141	140	139	138	140
11_Asahikawa	175	169	178	175	167	169	172	176	174	174	196	181	169	141	145	146	143	145	147
12_Nayoro	176	170	178	176	171	169	171	178	176	179	181	196	176	145	147	147	146	146	148
13_Soya	167	159	166	166	159	158	162	164	163	168	169	176	196	153	155	157	153	150	152
14_Ochiho	141	139	143	144	139	139	136	135	134	137	141	145	153	196	175	179	177	177	179
15_Tarantomari	140	133	146	146	135	140	135	138	137	141	145	147	155	175	196	187	177	179	179
16_Maoka	141	136	146	146	137	140	138	139	137	140	146	147	157	179	187	196	184	184	186
17_Shiraura	141	136	147	145	136	139	138	136	136	139	143	146	153	177	177	184	196	185	184
18_Raichishka	140	135	146	145	135	139	135	135	134	138	145	146	150	177	179	184	185	196	181
19_Nairo	144	137	149	149	138	143	138	139	138	140	147	148	152	179	179	186	184	181	196

subjective impression of the informants for one dialect to others, and are considered the significant reference for the classification of Ainu dialects. Or, rather, the scientific classification of Ainu dialects is considered to be unestablished. Therefore, this table can represent one attempt at it." Thus, Hattori and Chiri (1960) might consider Table 3 from "mutual intelligibility" in linguistics.

2.1 Automatically quantified “missing values” in lexicostatistics

This subsection illustrates that the four symbols (i.e., ○: “*questionable etymology or choice*”; ?: “*doubtful record*”; •: “*no answer given*”; (): “*lacuna of record.*”), which Hattori and Chiri exclude as “missing values” of either numerators or denominators from Table 4, are automatically quantified as some rational numbers in their calculations.

For example, the information shown on a map does not depend on whether the map is displayed in kilometers or meters as a unit. Therefore, the data in Table 4 were multiplied by 1.96 to normalize the total number of words in each dialect (i.e., the diagonal elements in Table 4) to 196 words (i.e., the diagonal elements in Table 6), as any positive and real constant by which Table 4 is multiplied does not affect the results of the classification applied to it[12]. The results are shown in Table 8.

Table 8. Part of the similarity data among the 19 Ainu dialects calculated by multiplying the raw data in Table 4 with 1.96 to set the total number of words to 196, rounded to the third decimal.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1_Yakumo	196.00	188.38	180.00	177.00	174.05	174.67	173.77	171.50	172.91	171.88	175.48	169.87	166.24
2_Oshamambe	188.38	196.00	179.75	176.70	177.43	176.40	176.09	170.61	172.27	169.97	173.59	167.06	162.98
3_Horobetsu	180.00	179.75	196.00	183.00	177.63	180.26	172.26	173.50	177.53	172.38	177.02	172.88	164.19
4_Biratori	177.00	176.70	183.00	196.00	187.32	188.38	168.22	170.00	172.40	169.36	173.94	170.87	163.68
5_Nukibetsu	174.05	177.43	177.63	187.32	196.00	187.70	166.60	165.89	168.00	164.19	171.89	168.44	162.46
6_Niikappu	174.67	176.40	180.26	188.38	187.70	196.00	171.24	168.95	172.42	168.58	173.06	166.91	161.07
7_Samani	173.77	176.09	172.26	168.22	166.60	171.24	196.00	177.72	180.53	174.78	175.26	167.06	165.41
8_Obihiro	171.50	170.61	173.50	170.00	165.89	168.95	177.72	196.00	185.68	182.43	176.40	173.77	164.89
9_Kushiro	172.91	172.27	177.53	172.40	168.00	172.42	180.53	185.68	196.00	184.20	179.14	173.94	169.13
10_Bihoro	171.88	169.97	172.38	169.36	164.19	168.58	174.78	182.43	184.20	196.00	174.85	173.27	168.66
11_Asahikawa	175.48	173.59	177.02	173.94	171.89	173.06	175.26	176.40	179.14	174.85	196.00	178.98	167.48
12_Nayoro	169.87	167.06	172.88	170.87	168.44	166.91	167.06	173.77	173.94	173.27	178.98	196.00	171.76
13_Soya	166.24	162.98	164.19	163.68	162.46	161.07	165.41	164.89	169.13	168.66	167.48	171.76	196.00

It was observed that the similarity between the Yakumo and Samani dialects (Nos. 1 and 7 in Figure 2) was 172.0 in Table 6 and about (172/194)*100*1.96≒173.77 in Table 8, and the similarity between the Samani and Nayoro dialects (Nos. 7 and 12 in Figure 2) was 164.5 in Table 6 and (164.5/193)*100*1.96≒167.06 in Table 8.

Two symbols (i.e., •: “*no answer given*”) are included between the Yakumo and Samani dialects and the same three symbols between the Samani and Nayoro dialects. Thus, the former quantified the symbol as (173.77-172)/2≒0.89 and the latter as (167.06-164.5)/3≒0.85 “automatically” as results.

Since Hattori and Chiri’s calculation method consisted of only three numbers (i.e., 0, 0.5, and 1), as discussed in Section 2.2, it is questionable from both linguistic and mathematical perspectives that •: “*no answer given*” is “automatically” assigned by a rational number and the assigned rational numbers are different in the same symbol.

[12] It is trivial that each dialect is cognate to itself in the 196 words.

Since each pair of dialects contains different amounts of the four symbols (i.e., ○, ?, •, and ()) and leads to different denominators in Table 5, these uninterpretable phenomena occurred on "missing values" in Hattori and Chiri's calculations.

Notably, most previous lexicostatistical studies (e.g., Black 1973, Dyen, Kruskal, and Black 1992) have calculated the similarity between dialects or languages as "lexicostatistical percentages" or "cognate percentages," resulting in the same critical problem in automatically quantifying the "missing values."

Therefore, this study did not adopt Hattori and Chiri's calculation method. Instead, the following subsection scrutinizes an alternative on the similarity in lexicostatistics from the viewpoint of the definition for the numbers. The discussion reveals Asai's calculation methods as the more appropriate approach.

2.2 How to define the numbers for language metrics: an algebraic problem

This subsection focuses on how to address the numbers in lexicostatistics from algebraic viewpoints. There are mainly two approaches to analyzing lexicostatistical data in Ainu dialects and classifying them. The first considers that the seven symbols in Hattori and Chiri indicate some uncertainty in cognacy judgments in the linguistic environment surrounding Ainu dialects and assigns these symbols as continuous values on an ordinal scale in terms of the uncertainty (Ono 2019a, 2019b). The second approach measures the similarity between Ainu dialects with respect to the remaining words, the basic concept in lexicostatistics or glottochronology, and quantifies the symbols as discretized non-negative integers (Hattori and Chiri 1960, Asai 1974b, Ono 2020a, 2020b). Since this paper focuses on lexicostatistical rules in the symbols, this section applies the latter method to Hattori and Chiri.

I summarized the definition of the numbers in Hattori and Chiri (1960), which assigns + as 1 when "there is at least one cognate word form in both of two given dialects, and there is no non-cognate word form in either of two given dialects on 1 word"; ± as 0.5: "there is at least one cognate word form in both of two given dialects, and there is at least a non-cognate word form in either of two given dialects on 1 word"; − as 0: "there is no cognate word form in two given dialects on 1 word."

The unit is fundamental to measure objects accurately, and researchers should avoid different definitions of the unit unless the definitions are equivalent to each other or can be justified as substantive knowledge. For example, if two scales always weighed differently, it would be difficult to conduct business "at least on earth."

However, multiplying Hattori and Chiri's definition of 0.5 by two, Hattori and Chiri's definition contradicts the original definition of 1: "there is at least one cognate word form in both of two given dialects, and there is at least a non-cognate word form in either of two given dialects on 2 words." Thus, their definition of numbers is not mathematically supported and lacks any justification from linguistics.

Asai defined + and ± as 1[13]: "there is at least one cognate word form in both of two given dialects on 1 word"; − as 0: "there is no cognate word form in two given dialects on 1 word." Since the unit of lexicostatistics is uniquely determined in Asai (1974b), this paper adopts Asai's definition as more admissible in lexicostatistics[14].

Section 4 shows that Hattori and Chiri's and Asai's similarity calculation consequently produce different classification of Ainu dialects.

3. Methods

This section discusses that Asai's (1974b) definition of the numbers contains only similarity information among Ainu dialects, and it is meaningless from a linguistic viewpoint for researchers to subtract the similarity data in Table 7 and utilize these values as "dissimilarity."[15]

As Ono (2020a: 28–29 [in Japanese]) and Ono (2020b: 238–239) have discussed, the logical interpretation of Asai's definition excludes the possibility of defining -1 without contradiction.

As the number -1 requires the condition that leads to the definition of 0, when researchers add +1 to -1, the tentative definition of -1 as "there is at least one cognate word form in both of two given dialects on -1 word" leads to the following definition of 0 as "there is at least one cognate word form in both of two given dialects on 0 word," contradicting the original definition of 0 as "there is no cognate word form in two given dialects on 1 word."[16]

Further, most traditional classification methods in statistics assume that the data contain both similarity and dissimilarity information and at least the order of similarity can be preserved as the order of dissimilarity and vice versa (e.g., inner scalar product model[17]). An alternative approach, which precludes the use of -1, should apply to Hattori and Chiri's similarity data.

Previous studies (Ono 2020a, 2020b) have proposed an alternative to spectral clustering, a graph-theoretic method that utilizes only similarity information from data to classify the dialects without requiring information on dissimilarity. Although various methods exist, such as the MinmaxCut algorithm (Mcut algorithm, hereinafter; Ding et

[13] The definition presupposes that there is at least one word form in both dialects. I simplify the definition of the numbers for the sake of explanation, based on the "relation index" in Asai (1974b: 61–62). Interested readers should therefore refer to Asai (1974b).

[14] Interested readers will observe that the definition of the numbers in both Hattori and Chiri (1960) and Asai (1974b) covered all possibilities about the judgment in two given dialects on 1 word.

[15] The same analysis can apply to Hattori and Chiri's definition of the numbers. Thus, Section 4 applies the statistical analysis adopted in Section 3 to Tables 3, 4, and 6.

[16] Interested readers should refer to Ono (2020b: 239; Footnote 8) in detail.

[17] In a rough sketch, inner product model corresponds to map on a plane. On the map, researchers can move two steps to the north and two steps to the east and still return to your initial position, if the map contains -1. Thus, inner product model does not hold in this case. cf. Ono (2022b, to appear).

al. 2001) and the NormalizedCut algorithm (Shi and Malik 2000), Ono (2022a) demonstrated that the Mcut algorithm possesses the desirable mathematical properties for the lexicostatistical classification of dialects or languages.

I, therefore, applied the Mcut algorithm to Table 7. Mcut algorithm aims 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; the division should maximize the sum of the similarity in the groups and minimize the sum of the similarity among the groups.

The advantages of the Mcut algorithm are summarized in three points. First, in the case of two divisions, the objective function of the Mcut algorithm is as follows: Mcut (A, A^c) = Cut (A, A^c) / W (A)+Cut (A^c, A) / W (A^c), where Cut (A, A^c) is the sum of the similarities between group A and the complement of group A, W(A) the sum of the similarities in group A, and W(A^c) the sum of the similarities in the complement of group A. Cut (A, A^c), W(A), and W(A^c) apply only the addition to the similarity data; in other words, this function does not require the negative numbers or the dissimilarity in the data, which is desirable for the properties in data discussed in this section.

Second, the Mcut algorithm can form a hierarchy of objects (e.g., a hierarchy of dialects, in my case) corresponding to the notion of hierarchical clustering methods, such as the complete linkage method or Ward's minimum variance method in statistics. Moreover, the Mcut algorithm searches A and A^c for a subdivision when a researcher needs more than three divisions.

Third, researchers can compute the objective function of Mcut (A, A^c) on all possible partitions and order all partitions in increasing sequence in terms of the values in Mcut (A, A^c). This property enables researchers to analyze the partition of dialects, whose objective function of the Mcut algorithm is not minimal in the given numbers of division but is important in the classification of dialects or languages.

As a previous study (Ono 2020a) observed ABA-type classification in Ainu dialects by calculating the objective function of Mcut algorithm on all partitions of Hokkaido Ainu dialects (Nos. 1–13 and 21 in Figure 2) with Asai's data, I also utilized the same approach with Hattori and Chiri, verifying the ABA-type classification in Ainu dialects statistically.

I utilized the Mcut algorithm implemented with R (R Core Team 2018) in Shinnou (2007: 132–133) in subsequent analyses.

4. Results

Figure 3 demonstrates the classification of Ainu dialects, applying the Mcut algorithm to Table 7. Figure 4 illustrates the result from Figure 3 on the map. Comparing these figures to Figure 2 using the same method as Asai's similarity data, I observed that the classification of Hokkaido Ainu dialects was generally consistent with that in Figure 2, whereas the Kushiro dialect (No. 9 in Figure 2) was farther to the Obihiro and Bihoro dialects (Nos. 8 and 10 in Figure 2) in Figures 3–4 than in Figure 2.

Furthermore, the classification of Sakhalin Ainu dialects (Nos. 14–19 in Figure 2) is more complicated. Figures 3–4 illustrate that the Sakhalin Ainu dialects are classified into an east coast group (Nos. 14, 17, and 19 in Figure 2) and a west coast group (Nos. 15, 16, and 18 in Figure 2), which is consistent with previous studies (Ono 2015, 2019a, 2019b), but Figure 2 does not show the same classification.

Furthermore, Figures 5–6 illustrate the classification results by Mcut algorithm applied to Tables 3–4, respectively. Figures 5–6 show that Hattori and Chiri's miscalculations cause different classification results, and Figures 3 and 6 demonstrate that Hattori and Chiri's and Asai's similarity calculations yield different classifications of Ainu dialects as a result.

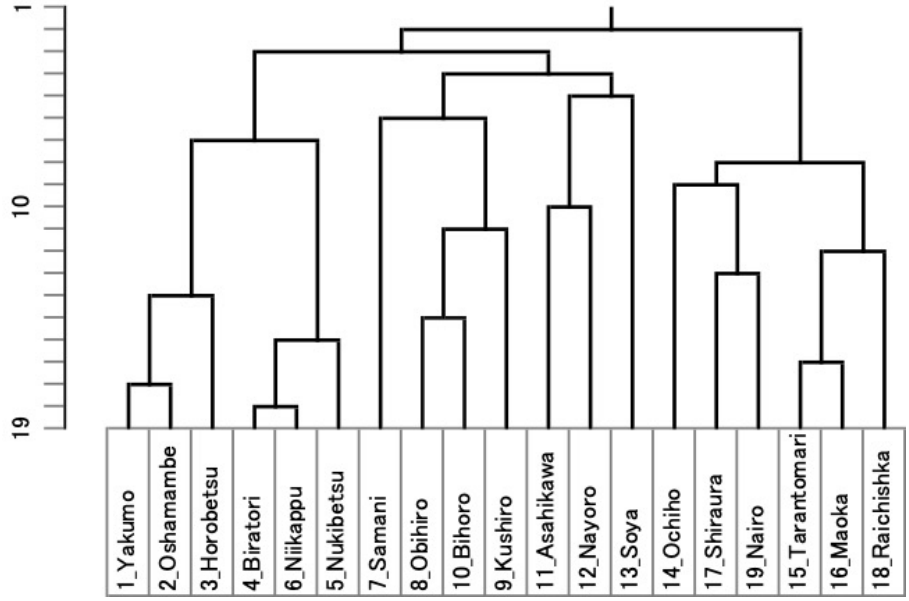

Figure 3. Result of the Mcut algorithm applied to Table 7 as a dendrogram.

As Section 2.2 concluded the similarity data in Table 4 as inadmissible from both linguistics and statistics viewpoints, the classification of the Sakhalin Ainu dialects should be "at least statistically" reconsidered so far as researchers recognize the classification of Sakhalin Ainu dialects obtained by Hattori and Chiri's similarity calculations.

Here, I report earlier comments on the classification of Sakhalin Ainu dialects related to the east coast group (Nos. 14, 17, and 19 in Figure 2) and the west coast group (Nos. 15, 16, and 18 in Figure 2)[18]. Therefore, selecting the appropriate classification of Sakhalin Ainu dialects from a linguistics perspective is best handled by Ainu linguists.

[18] Chiri (1955) stated: "In the Sakhalin Ainu dialects, there are some differences between the east and west coasts, but they are slight." Tamura (2000: 2) also stated that "*Sakhalin Ainu dialects are more similar to Hokkaido Ainu dialects than Kuril dialects are. There are differences between west coast and east coast dialects, but even so, these differences are not great. There are similar differences between northern and southern dialects.*"

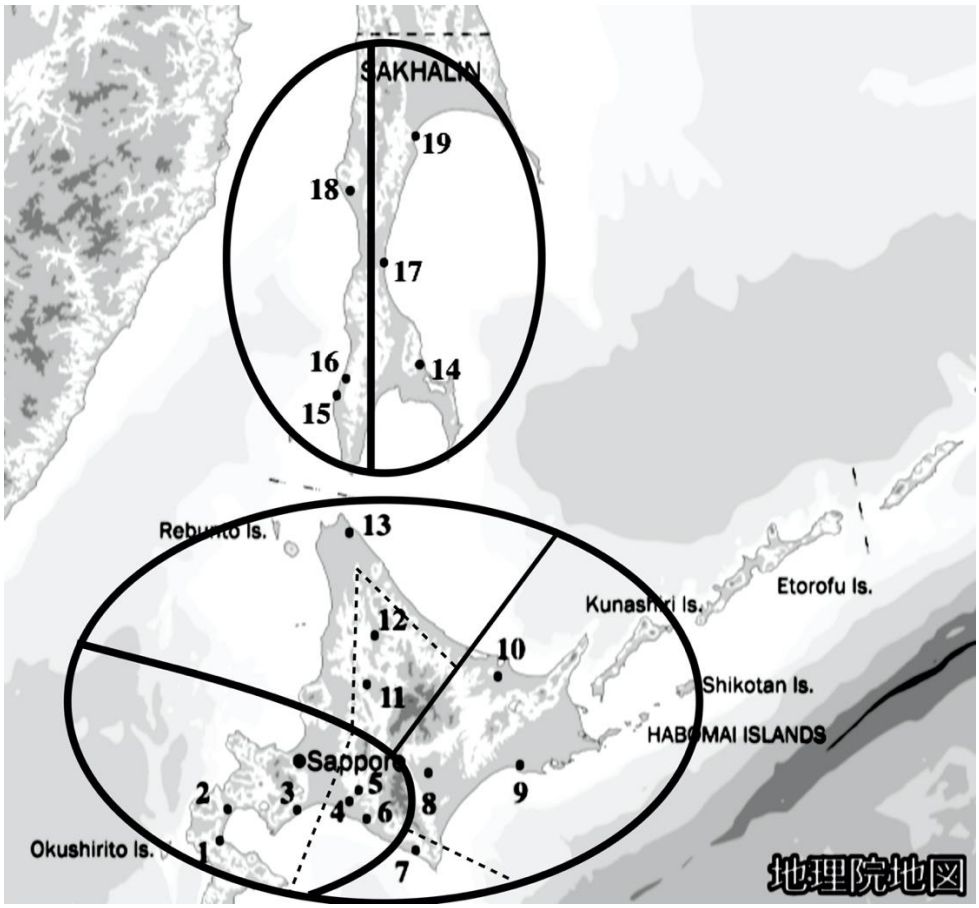

Figure 4. Map of the result in Figure 3 (Geospatial Information Authority of Japan, 2021), edited by the author. The number of divisions in the Mcut algorithm is 8.

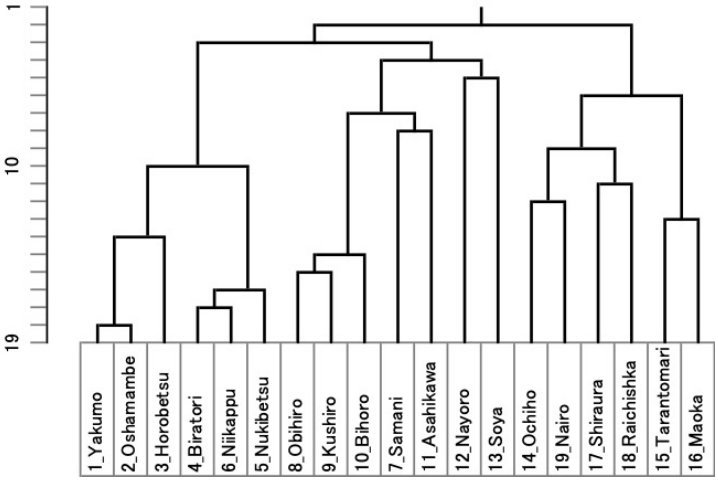

Figure 5. Result of the Mcut algorithm applied to Table 3 as a dendrogram.

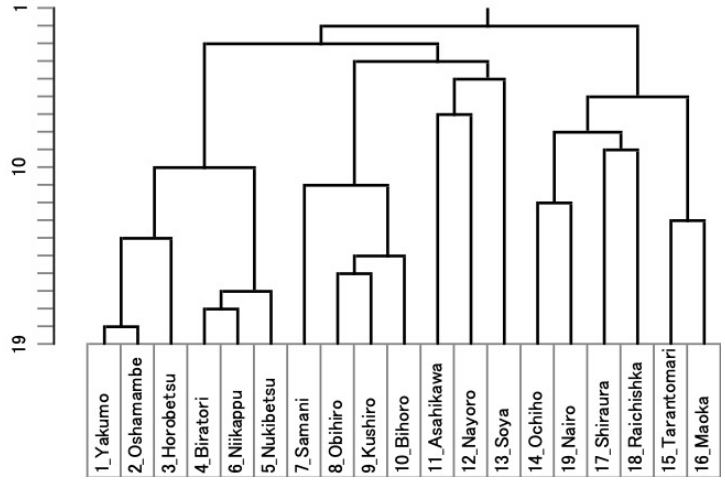

Figure 6. Result of the Mcut algorithm applied to Table 4 as a dendrogram.

Table 9 and Figure 7 demonstrate from the division under which the objective function of the Mcut algorithm is minimal to the division under which the objective function of the Mcut algorithm is the eighth minimal in the Hokkaido Ainu dialects (Nos. 1-13 in Figure 2). I summarize the characteristics in Table 9 and Figure 7 in the following five points: First, the Asahikawa dialect (No. 11 in Figure 2) belongs to the "southwestern" Hokkaido Ainu dialect group (Nos. 1–6 and 21 Figure 2) in the second division, while Asai (1974b: 100) classified the Asahikawa dialects in the "northeastern" Hokkaido Ainu dialect group (Nos. 7–13 in Figure 2). This result coincides with the reports in a previous study (Ono 2020a: 37–38; Figures 6–7).

Second, the Samani dialect belongs to the "southwestern" Hokkaido Ainu dialect group in the third division, which Asai (1974a) indicated by applying a different classification method to his data.

Table 9. Results of the Mcut algorithm ordered in increasing sequence (i.e., in the stronger order) on Hokkaido Ainu dialects (Nos. 1–13 in Figure 2). The numbers in groups A and A^c correspond to the number of places in Figure 2, and the value of Mcut (A, A^c) is rounded to the fourth decimal.

No. (division)	Group A	Group A^c	Mcut (A, A^c)
1	1, 2, 3, 4, 5, 6	7, 8, 9, 10, 11, 12, 13	3.235
2	1, 2, 3, 4, 5, 6, 11	7, 8, 9, 10, 12, 13	3.262
3	1, 2, 3, 4 ,5, 6, 7	8, 9, 10, 11, 12, 13	3.264
4	1, 2, 3, 4, 5, 6, 13	7, 8, 9, 10, 11, 12	3.266
5	1, 2, 3, 4, 5, 6, 12	7, 8, 9, 10, 11, 13	3.269
6	1, 7, 8, 9, 10, 12, 13	2, 3, 4, 5, 6, 11	3.279
7	1, 2, 4, 5, 6, 11	3, 7, 8, 9, 10, 12, 13	3.280
8	1, 8, 9, 10, 11, 12, 13	2, 3, 4, 5, 6, 7	3.282

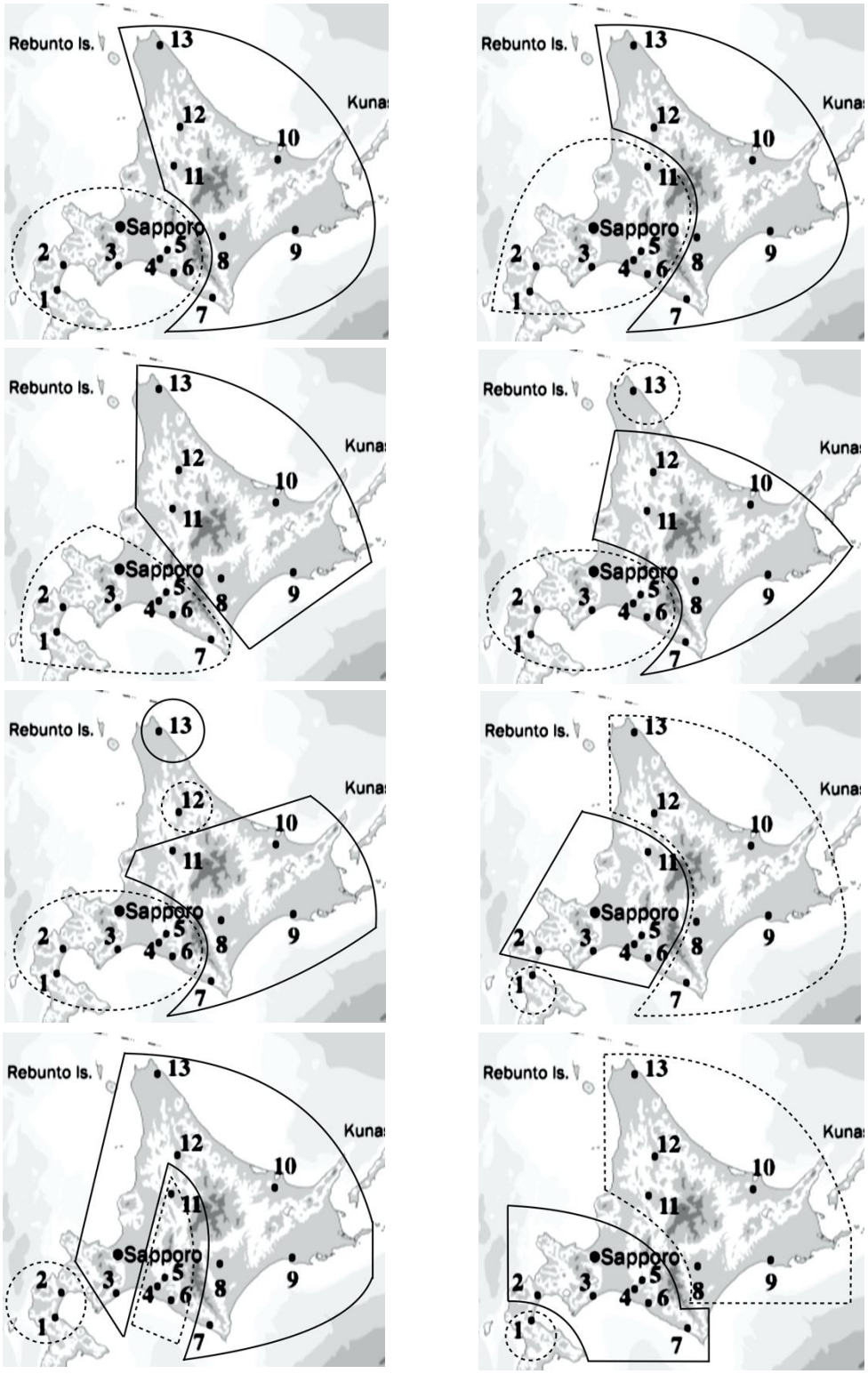

Figure 7. Map of the results in Table 9 (Geospatial Information Authority of Japan, 2021), edited by the author. Group A is surrounded by dotted lines and Group A^c is by solid lines, respectively. Top left: No.1 (in Table 9); Top right: No.2; Second top left: No.3: Second top right: No.4; Second bottom left: No.5; Second bottom right: No.6: Bottom left: No.7; Bottom right: No.8.

Third, I observed three types of ABA-type distribution: the fourth, sixth, and eighth divisions. In the fourth division, the Soya dialect (No. 13 in Figure 2) belongs to A in the ABA-type distribution; in the sixth division, the Yakumo dialect (No. 1 in Figure 2) also shows the same disposition in the ABA-type distribution; in the eighth division, the Samani dialect does not belong to B but to A in the ABA-type distribution, which will be important for current Ainu linguistics.

Fourth, the disposition of the Asahikawa and Nayoro dialects (Nos. 11–12 in Figure 2) is indeterminate in relation to the "southwestern" and "northwestern" groups and ABA-type distribution as shown in Ono (2019a), which will also matter for future research.

Finally, researchers need to pay more attention to the disposition of the Horobetsu dialect (No. 3 in Figure 2), which belongs to the "northeastern" Hokkaido Ainu dialect group in the seventh division.

5. Discussions and Conclusion

This section discusses the implications of the main results of this study for current and future linguistic research from both linguistics and statistics perspectives.

From the viewpoint of linguistics, the main results are summarized in four points. First, Asai's classification of the "northeastern" Hokkaido Ainu dialects, which clusters the Obihiro, Kushiro, Bihoro, Asahikawa, and Nayoro dialects (Nos. 8–12 in Figure 2), should be reconsidered so far as researchers recognize the classification based on Hattori and Chiri's and Asai's cognacy judgments. Both Figure 2, which applies the Mcut algorithm to Asai's cognacy judgments (Asai 1974b: 92; Table 1), and Figure 3, which applies the same method to Hattori and Chiri's cognacy judgments, are consistent in that the Obihiro, Kushiro, and Bihoro dialects form one cluster and the Asahikawa and Nayoro dialects form the other cluster, the latter of which is related to the Soya dialects (No. 13 in Figure 3).

Second, Table 9 and Figure 7 also demonstrated that the ABA-type distribution should be further investigated from the viewpoint of lexicostatistics in both Hattori and Chiri (1960) and Asai (1974b), as a previous study (Ono 2020a) demonstrated the significance of ABA-type distribution in Asai's lexicostatistical data via the same approach.

Third, researchers need to pay more attention to the classification of Sakhalin Ainu dialects. As Ono (2022c) discussed, their classification will play a significant role in examining whether the 91 basic words in Hattori and Chiri "statistically" transferred from Hokkaido to Sakhalin or from Sakhalin to Hokkaido.

Finally, the main results of this study suggest a reconsideration of the previous classification of Hokkaido Ainu dialects based on both Hattori and Chiri's and Asai's cognacy judgments (i.e., Figures 2–4). Thus, they were generally consistent with each other, except for the several differences discussed in Section 3.

From the viewpoint of statistics, the main results of this study indicate that previous lexicostatistical studies contain potentially two critical problems in the form of similarity data (i.e., Problem B in Figure 1).

The first problem concerns Hattori and Chiri's similarity calculation, in which the numerator (i.e., the number of cognate words) is divided by the denominator (i.e., the total number of words except the words whose cognacy judgment cannot be classified into cognate or non-cognate) as similarity between pairs of dialects (or languages), respectively. As illustrated in Section 2.1, the procedure "automatically" quantifies "missing values" in lexicostatistics as some uninterpretable rational numbers, resulting in the different classification shown in Figure 6.

The second problem concerns Hattori and Chiri's definition of numbers in lexicostatistics. As illustrated in Section 2.2, Hattori and Chiri's definition contains the contradiction in numbers. Thus, previous lexicostatistical studies need to be reconsidered (or recalculated) from the viewpoint of algebraic structure (i.e., by Asai's "relation index" method).

To the best of my knowledge, previous lexicostatistical research (e.g., Black 1973, Dyen, Kruskal, and Black 1992) calculated the similarity between pairs of dialects or languages with some methods, including Hattori and Chiri's two problems above. Given the above, I did not come across a study that calculated between pairs of dialects or languages as per Asai's method.

It is remarkable that, comparing Figure 3 and Figure 6, the two different calculation methods led to different results in the classification of dialects. Therefore, Asai's viewpoints in lexicostatistics encourage new developments in linguistics and dialectology, which will matter in future research[19].

I hope that the findings of this study will contribute to future developments in Ainu dialectology and that the new directions I propose lead future interdisciplinary research in the fields of humanities and information science/statistics in the form of an invaluable discipline[20].

---

[19] Recovering or examining the data omitted in previous studies owing to space restrictions will be challenging and promising research in the future, as this paper revised Tables 3 and 4, resulting in different results in Figures 5–6. Various symbols are generally used to record humanities data, and researchers must use considerable judgment when choosing among these symbols. However, such publications do not necessarily contain the full information on judgments, mainly because of space limitations prior to the advent of digital media or journals. Rather, the data were usually summarized as tables or figures, as in the case of Hattori and Chiri (1960) and Asai (1974b). Therefore, the complete information may not have been published and may be buried or lost in the researcher's library. However, as academic knowledge advances, it will be necessary to examine, compare, and integrate previous research in detail from current substantive viewpoints. This will require researchers to "recover" the complete information in previous research from the summarized format. Thus, the previous study (Ono 2020c) and this paper can be considered to be attempts to tackle this problem from a statistical approach.

[20] One of reviewers indicated the significance to coauthor with Ainu linguists for future research. I

Acknowledgments

I am very grateful to Prof. Megumi Kurebito (The University of Toyama) for her invaluable 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.

References

Asai, Tōru (1974a) Gengo kara mita chiiki shūdan [Regional groups from the viewpoints of the language]. In Niino, Naokichi and Hidezō Yamada (eds.), *Hoppō no kodai bunka [Ancient Culture in North Eurasian]*, 119-142, Tokyo: Mainichi-Shuppansha.

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

Benzécri, J.P. (1973) *L'Analyse des Donnénes. Vol.II: L'Analyse des Correspondances.* Paris: Dunod.

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

Bryant, David and Vincent Moulton (2004) Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2): 255-265.

Chiri, Mashiho (1953) *Bunrui ainugo jiten, Vol. I: Shokubutsu hen*. Tokyo: Nihon Jyōmin Bunka Kenkyūjo.

Chiri, Mashiho (1954) *Bunrui ainugo jiten, Vol. III: Ningen hen*. Tokyo: Nihon Jyōmin Bunka Kenkyūjo.

Chiri, Mashiho (1955) Ainu. *Sekai daihyakka jiten Vol.1 [Heibonsha World Encyclopedia Vol.1]*, 26-31. Tokyo: Heibonsha.

Crystal, David (ed.) (2011) *A dictionary of linguistics and phonetics 6th edition*. Oxford: Blackwell Publishing.

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

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

Fukazawa, Mika (2017) *Kagake monjo ni'okeru Ainugo no bunkengaku-teki kenkyū [A*

am currently collaborating with an Ainu linguist to improve cognacy judgments in the two previous studies, which will contribute to Ainu linguistics as lexicostatistical research consisting of advanced cognacy judgments and current statistical methodologies that statistician alone cannot achieve.

philological study of the Ainu language in the Kaga family's archives], Doctoral Thesis. Chiba: Chiba University.

Geospatial Information Authority of Japan (2021) Ministry of Land, Infrastructure, Transport and Tourism. URL: https://maps.gsi.go.jp [accessed on March 2021].

Gondran, Michel and Michel Minoux (2008) *Graphs, dioids and semirings: New models and algorithms.* New York: Springer Science+Business Media.

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

Hattori, Shirō (ed.) (1964) *Ainugo hōgen jiten [An Ainu dialect dictionary]*. Tokyo: Iwanami Shoten.

Hattori, Shirō (1967) Ainugo no oninkouzou to akusento: Ainusogo saikou no ichi kokoromi [Phonological structure and accent of Ainu: An attempt to reconstruct Proto-Ainu]. *Onsei no kenkyū [Study of sound]*, 13, 207-223. Tokyo: Phonetic Society of Japan.

Huson, Daniel and David Bryant (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(1): 254-267.

Kirikae, Hideo (1994) Pa/ca correspondence between Ainu dialects: A linguistic-geographical study. *The proceedings of the 8th international Abashiri symposium: Peoples and cultures of the boreal forest*, 8, 99-113. Abashiri: Hokkaido Museum of Northern People, Japan.

Lee, Alan and Bobby Willcox (2014) Minkowski generalizations of Ward's method in hierarchical clustering. *Journal of Classification*, 31(2): 194-218.

Murayama, Shichirō (1971) *Kita chishima ainu-go [Ainu language of Northern Kuril islands].* Tokyo: Yoshikawa-Kōbun-Kan.

Nakagawa, Hiroshi (1996) Gengo chirigaku ni yoru ainugo no shiteki kenkyū. [A historical study of the Ainu language through linguistic geography]. *Bulletin of the Hokkaido Ainu Culture Research Center*, 2, 1-17. Sapporo: Hokkaido Ainu Culture Research Center.

Ng, Andrew, Michael, Jordan and Yair Weiss (2002) On spectral clustering: Analysis and an algorithm. In Dietterich, Thomas, Suzanna Becker and Zoubin Ghahramani (eds.), *Advances in Neural Information Processing Systems*, 14, 849-856. Cambridge: MIT Press.

Ono, Yōhei (2015) Hattori and Chiri (1960) no toukei kagakutei saikousatsu: Ainugo hōgen shūken ron no jisshō [Statistical reanalysis of the classification of Ainu dialects: On the data of Hattori and Chiri (1960)]. *Hoppō jimbun kenkyū [Journal of the Center for Northern Humanities]*, 8: 25-41.

Ono, Yōhei (2019a) Observations on "Northeastern" Hokkaido Ainu dialects: A statistical perspective. *Hoppō gengo kenkyū [Northern Language Studies]*, 9: 95-

122.
Ono, Yōhei (2019b) The ordinal scale on lexicostatistical data in Ainu dialects: Towards a new interdisciplinary research among the humanities and statistics. *Hoppō jimbun kenkyū [Journal of the Center for Northern Humanities],* 12: 89-110.
Ono, Yōhei (2020a) Observations on lexicostatistical classifications on Hokkaido Ainu dialects: A new development on dialectology, graph-theoretic approach from algebraic structure. *Hokkaidō gengo bunka kenkyū [Journal of Language and Culture of Hokkaido],* 18: 19-46.
Ono, Yōhei (2020b) Reconsideration of "major division" of Ainu dialects: A statistical reanalysis of Asai (1974). *Hoppō gengo kenkyū [Northern Language Studies]*, 10: 231-254.
Ono, Yōhei (2020c) Some remarks on cognacy judgments of Ainu dialects: on Asai (1974). *Hoppō jimbun kenkyū [Journal of the Center for Northern Humanities]*, 13: 37-57.
Ono, Yōhei (2022a) Another look at data science in the humanities: Reconsideration of "invariance" in the classification. Manuscript in Preparation.
Ono, Yōhei (2022b, to appear) Jimbun gaku to jōhō no kako genzai soshite mirai [The past, present, and future of the humanities and information], *Hoppō gengo kenkyū [Northern Language Studies]*, 12.
Ono, Yōhei (2022c) On the relationships among Hokkaido Ainu dialects and Sakhalin Ainu dialects: A statistical observation. Manuscript in Preparation.
R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL: https://www.R-project.org/.
Shi, Jianbo and Jitendra Malik (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.
Swadesh, Morris (1955) Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 24: 121-137.
Székely, Gabor and Maria Rizzo (2005) Hierarchical clustering via joint between-within distance: extending Ward's minimum variance method. *Journal of Classification*, 22: 151-183.
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.
Tamura, Suzuko (2000) *The Ainu language (ICHEL linguistic studies 29)*. Tokyo: Sanseidō.
Torii, Ryūzō (1903) *Chishima Ainu [Kuril Ainu].* Tokyo: Yoshikawa-Kōbun-Kan.
Venables, William and Brian Ripley (2002) *Modern applied statistics with S. Fourth*

edition. New York: Springer
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.

Appendix

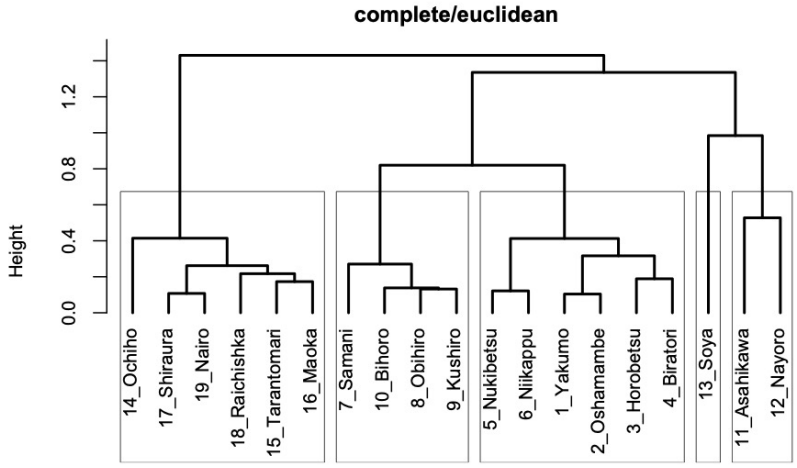

Figure 1- 1. The revised Figure 6 in Ono (2015: 33). Complete linkage method (Sørensen 1948) applied to the average 19 Euclidean distance matrix for each dialect, which was calculated by MASS package (Venables and Ripley 2002) in R language. For the sake of visualization, each of the five clusters is enclosed in a rectangle.

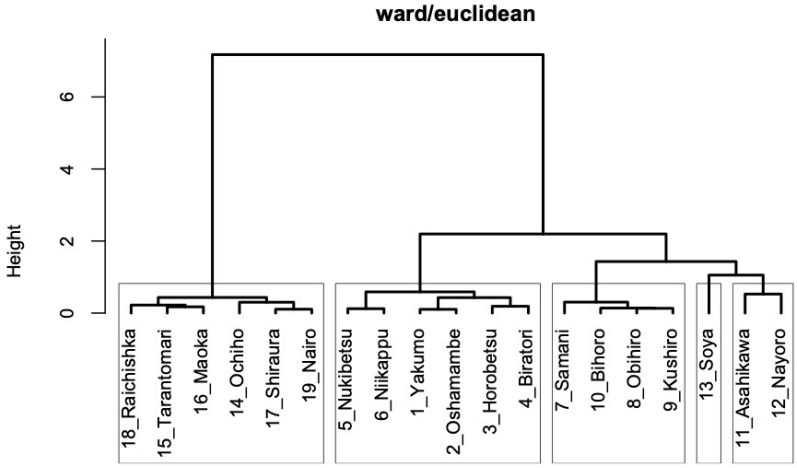

Figure 1- 2. The revised Figure 6 in Ono (2015: 33). Ward method (i.e., ε-method. See Ward 1963, Székeley and Rizzo 2005, Lee and Wilcox 2014) applied to the revised average 19 Euclidean distance matrix for each dialect. For the sake of visualization, each of the five clusters is enclosed in a rectangle.

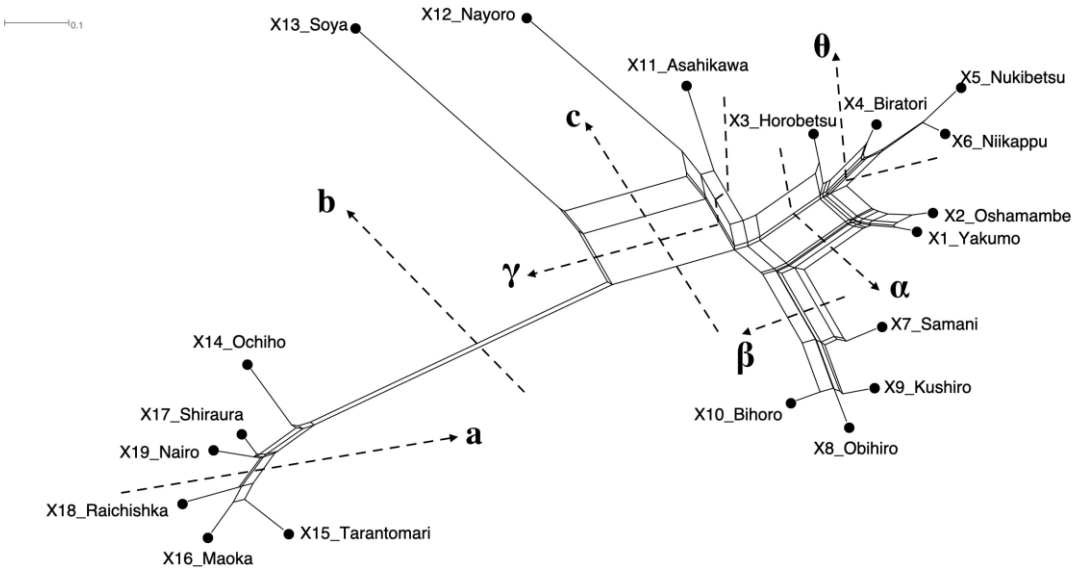

Figure 2- 1. The revised Figure 11 in Ono (2015: 36). Neighbor-Net (Bryant and Moulton 2004, Hudson and Bryant 2006) applied to the revised average 19 Euclidean distance matrix calculated by the three-dimensional coordinate obtained from the multiple correspondence analysis (Benzécri 1973) for each dialect. Lines a, b, and c correspond to those in Figure 11 in Ono (2015: 36), and lines, d, d', and e in the original figure do not hold in this figure. Lines α, β, γ, θ correspond to the "southwestern" group in the Hokkaido Ainu dialects and others, part of "northeastern" group in the Hokkaido Ainu dialects and others, the Asahikawa, Nayoro, and Soya dialects group and others, and the Saru-Chitose group and others, respectively. See Kirikae (1994), Nakagawa (1996), and Fukazawa (2017) for details.

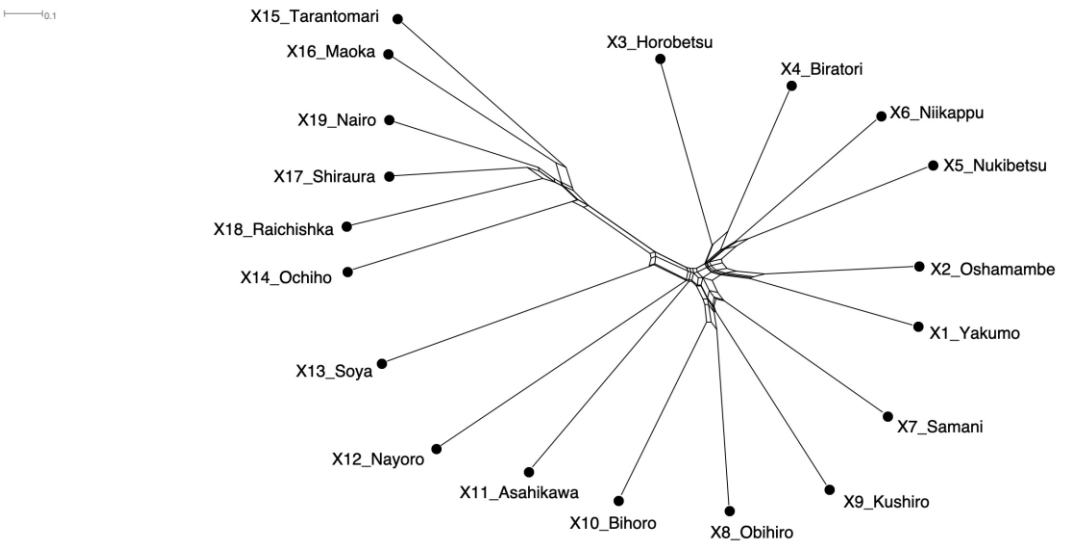

Figure 2- 2. Revised Figure 12 in Ono (2015: 36). Neighbor-Net (Bryant and Moulton 2004) applied to the revised average 19 Euclidean distance matrix by the 18-dimensional (i.e., all dimensions) coordinate obtained from the multiple correspondence analysis for each dialect.

Summary

There are two landmark lexicostatistical studies in Ainu dialectology focused on the classification of Ainu dialects, one by Hattori and Chiri (1960) and the other by Asai (1974b). However, two obstacles have prevented Ainu linguists from examining, comparing, and integrating Hattori and Chiri's and Asai's works. First, the two studies analyzed different lexicostatistical data and adopted different cognacy judgments. Second, the similarity among the Ainu dialects was measured using different methods in the two studies. This study focused on the second problem with the lexicostatistical data of Hattori and Chiri's cognacy judgments. The main results of this study are summarized as follows: First, I reported miscalculations in Hattori and Chiri and proposed a corrected similarity matrix. Second, I examined two critical problems in lexicostatistics: the automatic quantification on "missing values" and the definition of numbers for language metrics. The discussions unfolded an alternative in lexicostatistics from Asai's viewpoints. Third, I applied spectral clustering, a graph-theoretic approach adopted in previous studies (Ono 2020a, 2020b). Notably, the classification of Ainu dialects obtained statistically was generally consistent with those obtained by the same method to Asai's data. Several differences between the two classifications were also reported in the disposition of the Horobetsu, Samani, Kushiro, Asahikawa, and Nayoro dialects. Fourth, I investigated the other strongest classifications in the Hokkaido Ainu dialects. The result led to an ABA-type distribution and its variant, which a previous study (Ono 2020a) also reported for Asai's data using the same approach. The main results of this study suggest the need to reconsider not only Hattori and Chiri's and Asai's classification of Ainu dialects, but also previous lexicostatistical studies from methodological point of view.

（linguistics.dialectometry@gmail.com）