[Article]

The Ordinal Scale on Lexicostatistical Data in Ainu Dialects: Towards a New Interdisciplinary Research among the Humanities and Statistics

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

1. Introduction

Since a series of articles in the 1950's by Morris Swadesh addressed the relationship among dialects or languages as "lexicostatistics," his attempts have brought various insights regarding linguistics. The basic assumption of lexicostatistics is that linguists can determine the specific word form representing one word in a certain dialect (or language) and judge the cognacy among those word forms.

Hattori and Chiri's (1960) study, which we focus on in this paper, also comprises a lexicostatistical survey of 19 Ainu dialects (i.e., Nos. 1-19 in Figure 1) in the same period as Swadesh, referencing Swadesh's wordlist (Swadesh 1955). Therefore, we observe the typical data type of lexicostatistics in Hattori and Chiri (1960: 321) as shown in Table 1. Table 1 presents the original records of "rain" in 19 Ainu dialects and the binary data based on their cognacy judgments. Table 1 holds the basic assumptions of lexicostatistics that (1) linguists can determine the specific word form representing one word in a certain dialect (or language) and (2) judge the cognacy among those word forms. Note that the binary data in Table 1 are mutually exclusive. This corresponds to the basic assumption of lexicostatistics itself as explained above.

Figure 1. Map of a section of the region where the Ainu language is or was spoken (Geospatial Information Authority of Japan, 2018), 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.

At that time, Ainu dialects were on the verge of vanishing and the authors noted that "*Some of the informants were the last surviving speaker or speakers of dialects, and all of them were very old people. Some of them even have died since our investigation*" (Hattori and Chiri 1960: 307)[1)]. Therefore, Hattori and Chiri (1960) hold a position of monumental documentation in Ainu

1) In the following sentences, unless italicized, the English translation of Japanese literature is by the author.

research. Refsing (1986: 20) comments on their invaluable contribution to Ainu research, including Hattori and Chiri (1960), as having secured "*Ainu language research as a recognized and fairly respectable branch of linguistics in Japan.*"

Table 1. The records of "rain" in Hattori and Chiri (1960: 321) and the binary data according to their cognacy judgments. In each dialect, presence is coded as "1" and absence as "0."

	"rain" in Hattori and Chiri (1960)		Binary Data		
Dialect	Word Form	Dialect	wení-'as-type	'ápto-'as-type	ruyanpe-'as-type
X1_Yakumo	wení 'ás	X1_Yakumo	1	0	0
X2_Oshamambe	wení 'ás	X2_Oshamambe	1	0	0
X3_Horobetsu	'ápto 'as	X3_Horobetsu	0	1	0
X4_Biratori	ápto 'as	X4_Biratori	0	1	0
X5_Nukibetsu	ápto 'as	X5_Nukibetsu	0	1	0
X6_Niikappu	ápto 'as	X6_Niikappu	0	1	0
X7_Samani	ruyanpe 'as	X7_Samani	0	0	1
X8_Obihiro	ruyánpe rúy	X8_Obihiro	0	0	1
X9_Kushiro	ruwanpe 'as	X9_Kushiro	0	0	1
X10_Bihoro	ruwanpe 'as	X10_Bihoro	0	0	1
X11_Asahikawa	ruyánpe 'as	X11_Asahikawa	0	0	1
X12_Nayoro	ruyánpe 'as	X12_Nayoro	0	0	1
X13_Soya	ruyánpe 'as	X13_Soya	0	0	1
X14_Ochiho	'ahto raN	X14_Ochiho	0	1	0
X15_Tarantomari	'atto ran	X15_Tarantomari	0	1	0
X16_Maoka	'ahto ran	X16_Maoka	0	1	0
X17_Shiraura	'ahto ran	X17_Shiraura	0	1	0
X18_Raichishka	'ahto ran	X18_Raichishka	0	1	0
X19_Nairo	'atto ran	X19_Nairo	0	1	0

This paper addresses another pioneering aspect of their research that, despite the lexicostatistical survey at the beginning stage, they implied the limitations of lexicostatistics through studies of the endangered language for Ainu.

The first issue regards the data type of *n* dialects (or languages) and *m* words. For example, many lexicostatistical records found in Hattori and Chiri (1960) violate the basic assumption of lexicostatistics as shown in Table 2.

Table 2 shows the original records of "cloud" in 19 Ainu dialects and the cognacy judgment based on Yakumo dialect record according to their own linguistic knowledge.

Notably, Hattori and Chiri (1960) introduced a new symbol for cognacy judgments, "±," which represents the fact that "*one or both of the dialects have two or more forms, and imperfectness of the record does not allow researchers to decide which is more basic*" (Hattori and Chiri 1960: 307) and "the informant cannot exactly report the difference of usage among those (word) forms" (Hattori and Chiri 1960: 312). For example, the cognacy judgment between Yakumo and Kushiro is represented by "±" because the word forms in Kushiro dialect (i.e., 'urar and niskur), among which the Kushiro informant cannot precisely report the difference of usage, shares one word form, nískur/niskur or kúr, with Yakumo dialect.

However, the introduction of "±" to lexicostatistical data indicates that the basic assumption of lexicostatistics cannot hold. Since the original records in Kushiro dialect, represented as "'urar, niskur," mean that "*one or both of the dialects have two or more forms and imperfectness of the record does not allow*

researchers to decide which is more basic" (Hattori and Chiri 1960: 307), the notation of the records for Kushiro dialect violates the first assumption that linguists can determine the specific word form representing one word in a certain dialect (or language).

Furthermore, Hattori and Chiri (1960: 312) explained the notation of "±" in detail as "this symbol (i.e., '±') applies to cases in which there are two or more morphemes to compare in either dialect or both dialects. Normally, if the informant can report the difference in usage between the two (or more) word forms in detail and exactly, and the morpheme is correctly and exactly recorded as a result, it is clear which of the two (or more) word forms to select (as the basic word form). However, the case does not satisfy these conditions. Therefore, if the original records are represented by only two or more word forms, then we had no alternatives but to mark with the symbol: '±.' " This results in the violation of the second assumption of lexicostatistics that researchers can judge the cognacy among word forms[2)].

Table 2. The records of "cloud" in Hattori and Chiri (1960: 321) and the cognacy judgments data based on Yakumo dialect record according to their own linguistic knowledge.

	"cloud" in Hattori and Chiri (1960)	Cognacy Judgments based on Yakumo record in Hatttori and Chiri (1960)
Dialect	Word Form	Cognacy Judgments
X1_Yakumo	nískur, kúr	＋
X2_Oshamambe	kúr	＋
X3_Horobetsu	nískur	＋
X4_Biratori	nískur	＋
X5_Nukibetsu	nískur	＋
X6_Niikappu	nískur	＋
X7_Samani	nis	－
X8_Obihiro	nís, 'úrar	－
X9_Kushiro	'urar, niskur	±
X10_Bihoro	nis	－
X11_Asahikawa	nis, nískur, 'úrar	±
X12_Nayoro	nis, nískur	±
X13_Soya	'úrar, nískur	±
X14_Ochiho	niskuru	＋
X15_Tarantomari	'uurara	－
X16_Maoka	'uurara	－
X17_Shiraura	niskuru	＋
X18_Raichishka	niskuru	＋
X19_Nairo	niskuru	＋

(Note) ＋ : cognate residues, －: non-cognates, and ±: cognates and non-cognates, "*when one or both of the dialects have two forms, and the imperfectness of the record does not allow us to decide which is more basic*" (Hattori and Chiri 1960: 307).

Next, the significant question arises as to whether the lexicostatistical data violating the basic assumption of lexicostatistics can be represented as binary data for the application of statistical analysis as observed in

2) Note that the reinvestigation and further fieldwork normally compensate for the information to recover these violations but the lexicostatistical survey of the endangered language for Ainu does not necessarily provide these conditions. Consequently, Hattori and Chiri (1960) have focused on the fundamental problem of lexicostatistics in the beginning of this realm.

current research. Table 3 is a candidate for the alternative with binary attribute, which is desirable for many statistical analyses including Bayesian Phylogenetic Analysis (Lee and Hasegawa 2013).

Table 3. A candidate of binary data for the original records in Table 2. The original records are classified into three types: kúr–type, nis-type, and ’úrar-type. In each dialect, presence is coded as “1” and absence as “0.”

	A candidate for binary data		
Dialect	kúr-type	nis-type	'úrar-type
X1_Yakumo	1	0	0
X2_Oshamambe	1	0	0
X3_Horobetsu	1	0	0
X4_Biratori	1	0	0
X5_Nukibetsu	1	0	0
X6_Niikappu	1	0	0
X7_Samani	0	1	0
X8_Obihiro	0	1	1
X9_Kushiro	1	0	1
X10_Bihoro	0	1	0
X11_Asahikawa	1	1	1
X12_Nayoro	1	1	0
X13_Soya	1	0	1
X14_Ochiho	1	0	0
X15_Tarantomari	0	0	1
X16_Maoka	0	0	1
X17_Shiraura	1	0	0
X18_Raichishka	1	0	0
X19_Nairo	1	0	0

However, the logical interpretation of Table 3 shows that, for example, Kushiro dialect has kúr-type, which distinguishes its usage from nis-type and ’úrar-type, and ’úrar-type, which distinguishes its usage from kúr-type and nis-type. Therefore, Kushiro dialect has kúr-type and ’úrar-type, for both of which the Kushiro informant can distinguish their usage from each other but the notion of “cloud” in Japanese does not have these differences.

This result clearly contradicts the definition of the cognacy judgment, “±,” in Hattori and Chiri (1960) with respect to the two points. Again, note that “*one or both of the dialects have two or more forms and imperfectness of the record does not allow researchers to decide which is more basic*” (Hattori and Chiri 1960: 307) and “the informant cannot exactly report the difference of usage among those (word) forms” (Hattori and Chiri 1960: 312).

Furthermore, Hattori and Chiri (1960: 322) have marked the special notation (i.e., “{”) for the situation explained above in the case of “yellow” in Biratori dialect and Obihiro dialect. They have stated that “Since the notion of ‘yellow’ in Japanese (i.e., kiiroi) has both the meaning of ‘siwin’ and ‘hure’ in the dialect we use this special notation” (Hattori and Chiri 1960: 336).

Ono (2019) has compared various alternatives on the data type in Hattori and Chiri (1960), applying different statistical methods and evaluating the classification of Ainu dialect in terms of Ainu dialectology. Ono (2019) concluded that the data type in Hattori and Chiri (1960) produced the most consistent classification of 19 Ainu dialects and proposed an extension of data type for lexicostatistics. Therefore, the author will leave the problem of data type to Ono (2019) and deal with the main concern in this paper, assuming the same data as Hattori and Chiri (1960).

The other issue concerns the statistical methods used to analyze the data type of lexicostatistics in Hattori and Chiri (1960). A previous study (Ono 2015) has approached this problem, applying statistical methods (i.e., Multiple Correspondence Analysis, Neighbor-Net Analysis, and MCANeighbor-Net) to each cognacy judgment data corresponding to each dialect and calculating 19 distance matrices from these data. This procedure has enabled the whole relationship among 19 Ainu dialects to be visualized from the average

distance matrix of these 19 distance matrices.

Examples are shown in Table 4, using the data in Table 2. There are several statistical methods for calculating the distance from these cognacy judgment matrices. One way is to count the number of data as the distance between A dialect and B dialect, where A dialect shows "1" and B dialect shows "0" or A dialect shows "0" and B dialect shows "1." Another way is to quantify (i.e., to assign "some appropriate" numerical values from statistics) each of " ＋ ," "±," and "−" for each dialect respectively and to calculate the distance matrices based on these quantifications.

There are various statistical methods to quantify these data such as Dual Scaling (Nishisato 2006), Hayashi's Quantification Method Ⅲ (Hayashi 1952), Homogeneity Analysis (Gifi 1990), and Multiple Correspondence Analysis (Benzécri 1973). These methods are shown to be equivalent under certain mathematical conditions; the various names comprise a legacy of different applications to different fields of research.

Note that, in general, these quantification methods assume that the data to be quantified (i.e., " ＋ ," "±," and "−" in our case) are on a nominal scale.

The nominal scale (or categorical data) has a mathematical property, on which we cannot place any relations, including addition, subtraction, multiplication, division, or order relation. Moreover, on the nominal scale, there only exists the distinction among the data (i.e., counting).

However, in Table 4, the author hypothesized whether the cognacy judgments including " ＋ ," "±," and "−" have an order relation, since there seems to be an order relation among " ＋ ," "±," and "−" as " ＋ " > "±" > "−" or " ＋ " < "±" < "−" with respect to linguistics knowledge in Table 4.

Therefore, the main objective of this paper is to examine the assumption on the nominal scale in lexicostatistics and to seek the possibility that an assumption of order relation among the cognacy judgments in Hattori and Chiri (1960) could improve the statistical classification of 19 Ainu dialects[3]. The author will evaluate the validity on the ordinal scale, comparing the classification results of 19 Ainu dialects on the ordinal scale with those on the nominal scale in terms of Ainu dialectology. Thus, this paper focuses on a statistical issue; that is, how we can appropriately analyze the lexicostatistical data of Hattori and Chiri (1960) with the assumption of the ordinal scale.

The main result in this paper demonstrates that the lexicostatistical data contain significant information on

3) Note that Fumio Inoue (1942~), a Japanese linguist who has pioneered the attempt to introduce Hayashi's Quantification Method III to Japanese Linguistics, noticed the disadvantage of Hayashi's Quantification Method III and noted in Inoue (1985/8: 206) that "…However, Hayashi's Quantification Method III is not omnipotent. For example, (1) Hayashi's Quantification Method III deals with each word form as unrelated. Therefore, the two cases are equivalent for the calculation process of Hayashi's Quantification Method III; the one case, in which one word form and another are different in one consonant, and the other case, in which one word form and another are completely different." Inoue's suggestion precisely captures the possibility of the ordinal scale among word forms as explained above in the case of lexicostatistical data in Hattori and Chiri (1960). Inoue understands Hayashi's Quantification Method III in the following sentence: "But the difference in one consonant might have significance in a certain region. Therefore, we cannot quantify the differences among word forms a priori; rather, it supposes that we esteem the result of Hayashi's Quantification Method III." However, Inoue goes on to point out several obstacles of Hayashi's Quantification Method III. Nevertheless, computational and engineering difficulties have prevented the statistical analysis assuming the ordinal scale from being addressed, as implied in Inoue (1985/8) and Hattori and Chiri (1960).

Table 4. The examples of binary data for Yakumo dialect and Samani dialect based on the cognacy judgments from each word form of "cloud." Presence is coded as "1" and absence as "0."

	Cognacy Judgments based on Yakumo record in Hatttori and Chiri (1960)	Binary data of Yakumo dialect		
Dialect	Cognacy Judgments	+	±	−
X1_Yakumo	+	1	0	0
X2_Oshamambe	+	1	0	0
X3_Horobetsu	+	1	0	0
X4_Biratori	+	1	0	0
X5_Nukibetsu	+	1	0	0
X6_Niikappu	+	1	0	0
X7_Samani	−	0	0	1
X8_Obihiro	−	0	0	1
X9_Kushiro	±	0	1	0
X10_Bihoro	−	0	0	1
X11_Asahikawa	±	0	1	0
X12_Nayoro	±	0	1	0
X13_Soya	±	0	1	0
X14_Ochiho	+	1	0	0
X15_Tarantomari	−	0	0	1
X16_Maoka	−	0	0	1
X17_Shiraura	+	1	0	0
X18_Raichishka	+	1	0	0
X19_Nairo	+	1	0	0
	Cognacy Judgments based on Samani record in Hatttori and Chiri (1960)	Binary data of Samani dialect		
Dialect	Cognacy Judgments	+	±	−
X1_Yakumo	−	0	0	1
X2_Oshamambe	−	0	0	1
X3_Horobetsu	−	0	0	1
X4_Biratori	−	0	0	1
X5_Nukibetsu	−	0	0	1
X6_Niikappu	−	0	0	1
X7_Samani	+	1	0	0
X8_Obihiro	±	0	1	0
X9_Kushiro	−	0	0	1
X10_Bihoro	+	1	0	0
X11_Asahikawa	±	0	1	0
X12_Nayoro	±	0	1	0
X13_Soya	−	0	0	1
X14_Ochiho	−	0	0	1
X15_Tarantomari	−	0	0	1
X16_Maoka	−	0	0	1
X17_Shiraura	−	0	0	1
X18_Raichishka	−	0	0	1
X19_Nairo	−	0	0	1

the ordinal scale, which both linguistics and statistics can validate from their own substantive knowledge. Therefore, in the case of an endangered language like Ainu, the author recommends applying the ordinal scale to the lexicostatistical data, instead of regarding the data as a nominal scale: binary patterns without any ordinal relationship in the wordlist.

The remainder of this paper is organized as follows. Section 2 states that the lexicostatistical data in Hattori and Chiri (1960) are represented by various symbols that are beyond the scope of lexicostatistics, and indicates that we can consider these data as ordinal data from the viewpoint of linguistics. Furthermore, the

background on Homogeneity Analysis (Gifi 1990) with the assumption of the ordinal scale is introduced.

Section 3 demonstrates that statistical analyses on an ordinal scale more effectively visualize the relationship among 19 Ainu dialects consistent with Ainu dialectology. The author focuses on two dialects: Biratori and Samani. While several linguistic studies have noted the significance of these two dialects, statistical analyses (Hattori and Chiri 1960; Asai 1974; Lee and Hasegawa 2013) cannot visualize the uniqueness of these dialects based on the lexicostatistical data in Hattori and Chiri (1960). Our approach demonstrates that statistical analyses on the ordinal scale clearly illustrate the relationships in and around Biratori and Samani dialects compared to the nominal scale. The results reconfirm the philological knowledge in Ainu dialectology but are first verified statistically.

Section 4 discusses the significance of this paper. The assumption of scale type in lexicostatistical data is dealt with in linguistics and statistics[4]. Since almost all linguistic research applying statistical analysis to lexicostatistical data has considered the data as a nominal scale, the results in Section 3 lead us to reconsider previous lexicostatistical research from the statistical viewpoint. Furthermore, in general, statistical research in the humanities today has been polarized into two competing positions. The first school is typically represented by the Likert scale (Likert 1932) in social psychology, which imposes strong assumptions on the data: addition, subtraction, multiplication, and division (i.e., interval scale). The second school is represented by various quantification methods in psychometrics, which does not assume any relation among the data including addition, subtraction, multiplication, division, and order relation (i.e., nominal scale). The background of the polarization is discussed.

As a conclusion, this paper illustrates that complementary studies among linguistics and statistics will be promising for interdisciplinary research in the future.

2. Materials and Methods of Quantification and Evaluation

Section 2 focuses in detail on the issues in relation to lexicostatistical data in Hattori and Chiri (1960) and explains the statistical methods used to examine the hypothesis in this paper (i.e., the ordinal scale in lexicostatistics).

Section 2.1 focuses on four symbols: "○," "?," "•," and "(),"which are explained in Hattori and Chiri (1960: 307) as follows: "○": *questionable etymology or choice*, "?": *doubtful record*, "•": *no answer given*, and "()": *lacuna of record*. However, as demonstrated in Section 2.1, these symbols are not "mere lack of information on cognacy," but represent the cognacy judgments in Hattori and Chiri (1960) that the two word forms are neither "*cognate residues*" (i.e., " ＋ ") nor "*non-cognates*" (i.e., "−") but rather "*not non-cognates*"; that is, there is some information on uncertainty in the data. Therefore, this paper summarizes these four symbols as "△" (i.e., "*not non-cognates*") and hypothesizes an order relation among " ＋ ," "±," "−," and "△" as " ＋ "< "±" < " △ " < "−" or " ＋ " > "±" > " △ " > "−."

Section 2.2 introduces a statistical technique to quantify the symbols in Section 2.1, Homogeneity Analysis, which can analyze the lexicostatistical data in Hattori and Chiri (1960) with the assumption of both the

4) The assumption of the ordinal scale in Hattori and Chiri (1960) brings us an interesting and novel result for Asahikawa dialect, Kushiro dialect, Nayoro dialect, and Soya dialect from the linguistic viewpoint. Due to space limitations, the author will deal with this issue in another article.

nominal scale and ordinal scale. Therefore, Homogeneity Analysis is appropriate for the validation of the hypothesis in this paper[5].

Section 2.3 demonstrates that the author evaluates the assumption of the ordinal scale in the lexicostatistical data with respect to Biratori dialect and Samani dialect. Since, in spite of the significance and uniqueness in and around the two dialects, the statistical analyses based on the data in Hattori and Chiri (1960) have not clearly visualized these linguistic relationships, the author determined Biratori dialect and Samani dialect as the criteria of evaluation for the assumption of the ordinal scale in lexicostatistical data.

2.1 Materials

This Section focuses on the four symbols in Hattori and Chiri (1960): " ○ ," "?," " • ," and "()." First, Table 5 shows the word forms of "leaf" in Hattori and Chiri (1960: 316) and the cognacy judgments data based on Yakumo dialect record (i.e., hám) according to their own linguistic knowledge.

Although Hattori and Chiri (1960: 307) explained " ○ " as "*questionable etymology or choice*," they also stated that "the /yam/-type in Sakhalin dialects might appear from /niiyam/-type, whose change can describe / niiham/-type to /niiyam/-type (i.e., leaves). If we can agree with this assumption, then the cognacy judgments between /hám/-type and /yam-type/ is ' ＋.' But we determined ' ○ ' as questionable etymology" (1960: 336).

Second, Table 6 shows the word forms of "mother" in Hattori and Chiri (1960: 316) and the cognacy judgments data based on Yakumo dialect record (i.e., hápo) according to their own linguistic knowledge. Although Hattori and Chiri (1960: 307) explained "?" as comprising a "*doubtful record*," they also stated

Table 5. An example of " ○ " in Hattori and Chiri (1960). The records of "leaf" in Hattori and Chiri (1960: 316) and the cognacy judgments data based on Yakumo dialect record according to their own linguistic knowledge.

	"leaf" in Hattori and Chiri (1960)		Cognacy Judgments based on Yakumo record in Hattori and Chiri (1960)
Dialect	Word Form	Dialect	Cognacy Judgments
X1_Yakumo	hám	X1_Yakumo	＋
X2_Oshamambe	hám	X2_Oshamambe	＋
X3_Horobetsu	hám	X3_Horobetsu	＋
X4_Biratori	hám	X4_Biratori	＋
X5_Nukibetsu	hám	X5_Nukibetsu	＋
X6_Niikappu	hám	X6_Niikappu	＋
X7_Samani	hám	X7_Samani	＋
X8_Obihiro	hám	X8_Obihiro	＋
X9_Kushiro	hám	X9_Kushiro	＋
X10_Bihoro	hám	X10_Bihoro	＋
X11_Asahikawa	hám	X11_Asahikawa	＋
X12_Nayoro	hám	X12_Nayoro	＋
X13_Soya	hám	X13_Soya	＋
X14_Ochiho	yaN,-m	X14_Ochiho	○
X15_Tarantomari	yam	X15_Tarantomari	○
X16_Maoka	yam	X16_Maoka	○
X17_Shiraura	yam	X17_Shiraura	○
X18_Raichishka	yam	X18_Raichishka	○
X19_Nairo	yam	X19_Nairo	○

5) See Ono (2015) for statistical details for cluster analysis and Neighbor-Net Analysis (Huson and Bryant 2006) utilized in Section 3.

that “The original records in ‘mother’ can be translated as ‘Okaasan’ (familiar name of mother in Japanese). But the informant in Ochiho dialect reported ‘mother’ as ‘unu,’ whose word form we found in other dialects as ‘Hahaoya’ (formal name of mother in Japanese). Therefore we determine the cognacy judgments among Ochiho dialect and the other dialects as neither ‘−’ nor‘＋’, rather as ‘?’” (1960: 337)[6].

Third, Table 7 shows the word forms of “straight” in Hattori and Chiri (1960: 331) and the cognacy judgments data based on Yakumo dialect record (i.e., ’íkne and ’o’upéka) according to their own linguistic knowledge.

Although Hattori and Chiri (1960: 307) explained “ • ” as “*no answer given*,” they also stated that “The informant of Soya dialect answered ‘straight’ as /réwke somóki/ (i.e., not turn) and, he said, as neither /’o’upeka/-type nor /ku’anno/-type. If the investigator asked /’ittusne/-type, the informant of Soya dialect could remember. Therefore, we determined the cognacy judgments as ‘ • ’” (1960: 337).

Finally, Table 8 shows the word forms of “here” in Hattori and Chiri (1960: 324) and the cognacy judgments data based on Yakumo dialect record (i.e., téta) according to their own linguistic knowledge. Although Hattori and Chiri (1960: 307) explained “()” as “*lacuna of record*,” they also stated that “All word

Table 6. An example of “?” in Hattori and Chiri (1960). The records are “mother” in Hattori and Chiri (1960: 328) and the cognacy judgments data based on Yakumo dialect record according to their own linguistic knowledge.

	"mother" in Hattori and Chiri (1960)		Cognacy Judgments based on Yakumo record in Hattori and Chiri (1960)
Dialect	Word Form	Dialect	Cognacy Judgments
X1_Yakumo	hápo	X1_Yakumo	＋
X2_Oshamambe	hápo	X2_Oshamambe	＋
X3_Horobetsu	hápo	X3_Horobetsu	＋
X4_Biratori	hápo	X4_Biratori	＋
X5_Nukibetsu	hápo	X5_Nukibetsu	＋
X6_Niikappu	hápo	X6_Niikappu	＋
X7_Samani	hápo	X7_Samani	＋
X8_Obihiro	hápo	X8_Obihiro	＋
X9_Kushiro	hápo	X9_Kushiro	＋
X10_Bihoro	hápo	X10_Bihoro	＋
X11_Asahikawa	tótto	X11_Asahikawa	−
X12_Nayoro	tótto	X12_Nayoro	−
X13_Soya	hápo	X13_Soya	＋
X14_Ochiho	unu	X14_Ochiho	?
X15_Tarantomari	nanna	X15_Tarantomari	−
X16_Maoka	nanna	X16_Maoka	−
X17_Shiraura	nanna	X17_Shiraura	−
X18_Raichishka	'onmo	X18_Raichishka	−
X19_Nairo	nanna	X19_Nairo	−

6) One of reviewers indicated that “unu” in Ochiho dialect could appear from syllabemes or syllable contraction of “un-nu” that is derived from loss of initial consonant phonemes or metanalysis of “nunnu,” vocalic alternation of “nanna.” If linguist can agree with these assumptions, then the cognacy judgments between “unu” and the others will change. Since the research on Ainu language has made a great advance from Hattori and Chiri (1960), the systematic review of the cognacy judgments in Hattori and Chiri (1960) from the perspective of current Ainu linguistics will be promising in future research. Moreover, the author hopes that the statistical advancement (e.g., Homogeneity Analysis with the assumption of an ordinal scale in this paper) will also be constructive.

forms except Niikappu dialect corresponded to 'koko-ni' in Japanese (i.e., koko [here] + ni [particle]) but 'té'or' in Niikappu dialect corresponded to only 'koko' in Japanese. Although 'te' in 'té'or' and 'te' in 'téta' must be the same morpheme, Bihoro dialect distinguished 'té'or' as 'koko' in Japanese and 'ta'ánta' as 'koko-ni' in Japanese. Therefore, there is no guarantee that 'té'or' in Niikappu dialect means 'koko-ni' in Japanese. We determine the cognacy judgments as '()'" (1960: 336)[7].

We observe that none of the four symbols (i.e., " ○ ," "?," " • ," and "()") represent "mere lack of information"; rather, they contain some information among the word forms based on the substantive linguistics knowledge in Hattori and Chiri (1960).

This paper proposes to summarize these four symbols as "△,"which means "not non-cognates" or "researcher does not necessarily decide the cognacy judgments between two word forms as '−' (i.e., non-cognates)." This proposal leads us to hypothesize an order relation among " ＋ ," "±," "−," and " △ " as " ＋ "< "±" < " △ " < "−" or " ＋ " > "±" > " △ " > "−."

Therefore, the author introduces statistical techniques to validate this hypothesis in the next Section: Homogeneity Analysis, a statistical method to quantify these symbols with the assumption of the nominal scale or ordinal scale[8].

Table 7. An example of " • " in Hattori and Chiri (1960). The records are "straight" in Hattori and Chiri (1960: 331) and the cognacy judgments data based on Yakumo dialect record according to their own linguistic knowledge.

	"straight" in Hattori and Chiri (1960)		Cognacy Judgments based on Yakumo record in Hattori and Chiri (1960)
Dialect	Word Form	Dialect	Cognacy Judgments
X1_Yakumo	'íkne, 'o'upéka	X1_Yakumo	＋
X2_Oshamambe	'o'upéka	X2_Oshamambe	±
X3_Horobetsu	'o'úpeka	X3_Horobetsu	±
X4_Biratori	'ówpeka	X4_Biratori	±
X5_Nukibetsu	'o'ópeka	X5_Nukibetsu	±
X6_Niikappu	'ówpeka	X6_Niikappu	±
X7_Samani	'ówpeka	X7_Samani	±
X8_Obihiro	'o'úpeka	X8_Obihiro	±
X9_Kushiro	'owpeka	X9_Kushiro	±
X10_Bihoro	'o'upeka	X10_Bihoro	±
X11_Asahikawa	'íttusne	X11_Asahikawa	±
X12_Nayoro	'o'úpeka	X12_Nayoro	±
X13_Soya	(réwke somóki)	X13_Soya	•
X14_Ochiho	ku'aNno, 'ukuruhne	X14_Ochiho	−
X15_Tarantomari	ku'anno	X15_Tarantomari	−
X16_Maoka	ku'anno, 'istusne	X16_Maoka	±
X17_Shiraura	'ikuruhne	X17_Shiraura	−
X18_Raichishka	e'iku'anno	X18_Raichishka	−
X19_Nairo	'o'ihtusno	X19_Nairo	±

7) Note that "koko" is the demonstrative pronoun and "koko-ni" is the adverb in Japanese. Therefore, "koko-ni" can correspond to "to this place," "in this place," "from that place" etc. in English.

8) The author notes at the end of this Section 2.1 that there may be a question regarding the empirical adequacy of summarizing all four symbols (i.e., " ○ ," "?," " • ," and "().") as one symbol (i.e., " △ "). This paper does not necessarily exclude another option to summarize the four symbols (e.g., two or three symbols). However, in general, there is a trade-off between micro-classification and summarization and some "optimal" points between micro-

Table 8. An example of "()" in Hattori and Chiri (1960). The records are "here" in Hattori and Chiri (1960: 324) and the cognacy judgments data based on Yakumo dialect record according to their own linguistic knowledge.

	"here" in Hattori and Chiri (1960)		Cognacy Judgments based on Yakumo record in Hattori and Chiri (1960)
Dialect	Word Form	Dialect	Cognacy Judgments
X1_Yakumo	téta	X1_Yakumo	+
X2_Oshamambe	téta	X2_Oshamambe	+
X3_Horobetsu	téta	X3_Horobetsu	+
X4_Biratori	téta	X4_Biratori	+
X5_Nukibetsu	téta	X5_Nukibetsu	+
X6_Niikappu	té'or	X6_Niikappu	( )
X7_Samani	ta'anta	X7_Samani	−
X8_Obihiro	ta'ánta	X8_Obihiro	−
X9_Kushiro	tanta	X9_Kushiro	−
X10_Bihoro	temanta	X10_Bihoro	○
X11_Asahikawa	téta	X11_Asahikawa	+
X12_Nayoro	téta, tánta	X12_Nayoro	±
X13_Soya	téta	X13_Soya	+
X14_Ochiho	teeta	X14_Ochiho	+
X15_Tarantomari	teeta	X15_Tarantomari	+
X16_Maoka	teeta	X16_Maoka	+
X17_Shiraura	teeta	X17_Shiraura	+
X18_Raichishka	teeta	X18_Raichishka	+
X19_Nairo	teyta	X19_Nairo	+

2.2 Methods of Quantification

Since, in general, linguistic data are recorded with many "symbols," linguistic research based on these symbols requires some quantification methods (i.e., to assign appropriate values to these symbols) in practice. Homogeneity analysis is a type of these quantification techniques equivalent to Hayashi's Quantification Method Ⅲ under certain mathematical conditions.

It is known that various quantification methods such as Homogeneity Analysis or Hayashi's Quantification Method are formularized as the eigenvalue problem in statistics and can be solved as certain numerical values. However, if we impose some ordinal restriction on the results of quantification methods as an example, the calculation process to obtain the numerical values satisfying both the eigenvalue problem and ordinal restriction is rather complex.

Imagine that a researcher solves the eigenvalue problem and obtains certain numerical values. However, the values do not necessarily (or not always) satisfy the condition that each value corresponding to each symbol in each word is in decreasing (or increasing) order (i.e., ordinal scale). Conversely, imagine that a researcher discovers numerical values that satisfy the condition that each value corresponding to each symbol in each word is in decreasing (or increasing) order (i.e., ordinal scale). However, the values are not necessarily the solution of the eigenvalue problem. Therefore, the researcher must find the numerical values, which satisfy

classification and summarization proposed in statistics (Sakamoto, Ishiguro, and Kitagawa 1983). Therefore, this paper tentatively summarizes the four symbols as one but the linguistic research including lexicostatistical survey for another endangered language need to consider this problem of summarization in accordance with the characteristics of the data (i.e., the number of dialects [or languages], words, and symbols).

the eigenvalue problem and ordinal restriction simultaneously.

However, another problem soon arises with regard to how to find (or calculate) the "best" solution in practice. Since there exist many solutions that simultaneously satisfy the eigenvalue problem and ordinal restriction "to some degree," the researcher should find (or calculate) the "best" solution without any maps or any loads. For example, suppose that we have lexicostatistical data consisting of *n* dialects and 50 words, for each of which there are approximately four word forms (or symbols in our case) that are imposed on the ordinal scale (i.e., in decreasing or increasing order) in each word. A rough sketch shows that the researcher must search approximately 200 dimensional spaces (4 word forms×50 words) for the "best" solution. We cannot grasp and search such higher dimensional space as intuitively as we walk around with a map in our three-dimensional space. Thus, we shall search the higher dimensional space with the map of mathematics and computational engineering instead of the real map.

The key idea in using mathematics as the map in this high dimensional space and finding (or calculating) the numerical values with the eigenvalue problem and ordinal restriction is that the researcher constructs numerical procedures, under which the second solution calculated by the first solution becomes closer to the "best" solution (i.e., the solution is the sequence increasingly close to the best solution) and the final solution leads to the "best" solution from any first solution we select (e.g., if there are two peaks in higher dimensional space, whether we reach the highest peak [the "best" solution] completely depends on the choice of loads) with the guarantee of mathematics.

The recent advancement in computational science has enabled us to tackle these problems. R (2018) language implements Homogeneity Analysis based on both the nominal scale and ordinal scale in "homals" package (de Leeuw and Mair 2009). Therefore, the author utilizes "homals" package in the following analysis.

2.3 Methods of Evaluation

This paper evaluates the validity of the assumption for the ordinal scale in the lexicostatistical data of Hattori and Chiri (1960), focusing on the data of two dialects: Biratori and Samani. First, in and around Biratori dialect, many linguistic studies (See Tamura 1970; Tamura 2000) have ascertained its uniqueness to other dialects but no recent studies apart from Ono (2015) have verified this statistically. Therefore, the author will demonstrate the unique characteristics in and around Biratori dialect more clearly than previous studies in terms of statistics.

Second, researchers have also pointed out the importance and distinct characteristics of Samani dialect (Hattori and Chiri 1960; Asai 1974; Sato 2002), while the uniqueness of Samani dialect to the other dialects in Hokkaido has not been demonstrated clearly in the manner of statistics.

In the next Section, the author will demonstrate that the statistical analysis with the assumption of the nominal scale has obscured the uniqueness of Biratori dialect and Samani dialect; in other words, the assumption of the ordinal scale, in which the information of some uncertainty in the original records of Hattori and Chiri (1960) are quantified appropriately, has clarified the uniqueness of these two dialects, which was indicated many times philologically but not statistically.

3. Results

Figures 2 and 3 comprise the results of Neighbor-Net Analysis applied to three-dimensional Euclidean distance matrices calculated for Biratori dialect, which Homogeneity Analysis yielded on the assumption of the ordinal scale and nominal scale respectively in Hattori and Chiri (1960). From the viewpoint of statistics, Figure 2, which assumes the ordinal scale in the data, shows less ambiguous structures (i.e., a number of shorter sides in Neighbor-Net corresponds to the ambiguity for fit in Neighbor-Net Analysis) rather than Figure 3, which assumes the nominal scale in the data. This suggests that the assumption of the ordinal scale succeeds in capturing the underlying information structure in Hattori and Chiri (1960). Conversely, the underlying information structure in Hattori and Chiri (1960) has been obscured by the assumption of the nominal scale in previous studies.

Furthermore, the bipartite Figure 2 by line A and the bipartite Figure 3 by line A' correspond to the Saru-Chitose-type in Nakagawa (1996: 11) and the bipartite Figure 2 by line B and the bipartite Figure 3 by line B' correspond to the Eastern-Western-type in Nakagawa (1996: 6). Since the side of the bipartite by A is longer than that by B in Figure 2, Figure 2 clearly demonstrates that the Saru-Chitose-type is the stronger structure among 19 Ainu dialects than the Eastern-Western-type from the viewpoints of Biratori dialect.

The other analyses support this finding. Biratori dialect and the other two located around Biratori (i.e., Nukibetsu dialect and Niikappu dialect) are more closely clustered and farther located from other groups, particularly Yakumo dialect, Oshamambe dialect, and Horobetsu dialect in the two three-dimensional plots on the top of Figure 4, which correspond to the Euclidean coordinates in Figure 2, than in the two three-dimensional plots on the bottom of Figure 4, which correspond to the Euclidean coordinates in Figure 3. Furthermore, the result of cluster analysis on the left in Figure 5 (i.e., on the ordinal scale) demonstrates the Saru-Chitose-type as a dendrogram in contrast to the Eastern-Western-type on the right in Figure 5 (i.e., on the nominal scale).

The same analyses can apply to the case of Samani dialect. Figure 6, which assumes the ordinal scale in the data, shows less ambiguous structures than Figure 7, which assumes the nominal scale in the data. Samani dialect in Figure 6 has (1) a longer side from the Net, which means the strongest originality in Samani dialect in Hokkaido Ainu dialect, and belongs to (2) the bipartite Figure 6 and Figure 7 by line C and C' respectively, that is, northeastern Hokkaido Ainu dialect group, (3) those by line D and D' respectively, that is, southwestern Hokkaido Ainu dialect group and Sakhalin Ainu dialect group, and (4) the bipartite Figure 6 by line E, that is, Sakhalin Ainu dialect group and Soya dialect. In particular, the final fact could be of concern for Ainu linguists. Furthermore, Samani dialect in Figure 6 shows a more original disposition than in Figure 7. Neighbor-Net Analysis in Figure 6 indicates that Samani dialect has weaker relationships among northeastern Hokkaido Ainu dialect group, southwestern Hokkaido Ainu dialect group, and Sakhalin Ainu dialect group (and Soya dialect) than that in Figure 7, in which the Samani dialect is closely located to the northeastern Hokkaido Ainu dialect group with the bipartite by line C'.

The other analyses also support this finding. Samani dialect is farther located from the other groups (i.e., northeastern Hokkaido Ainu dialect group, southwestern Hokkaido Ainu dialect group, and Sakhalin Ainu dialect group [and Soya dialect]) in the two three-dimensional plots on the top of Figure 8, which correspond to the Euclidean coordinates in Figure 6, than in the two three-dimensional plots on the bottom of Figure 8,

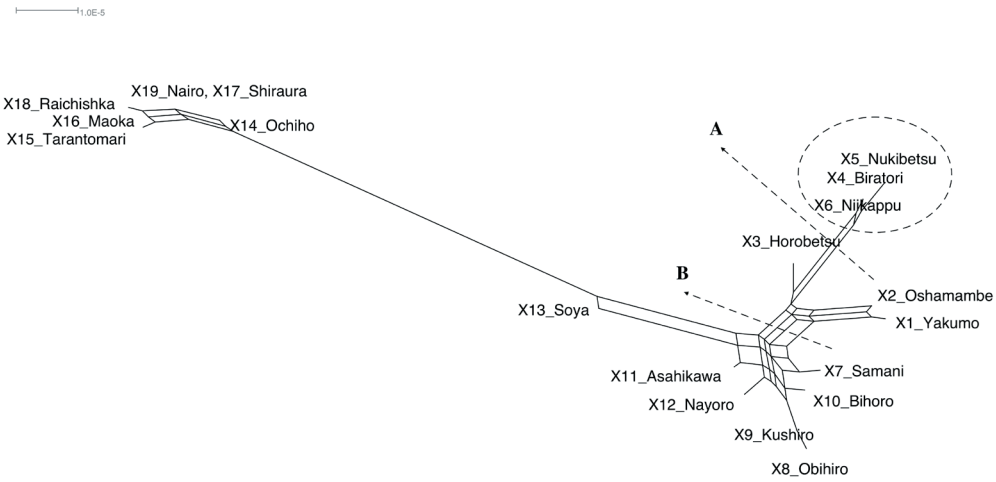

Figure 2. The result of Neighbor-Net Analysis applied to the Euclidean distance matrix for Biratori dialect of Homogeneity Analysis assuming the ordinal scale in Hattori and Chiri (1960)

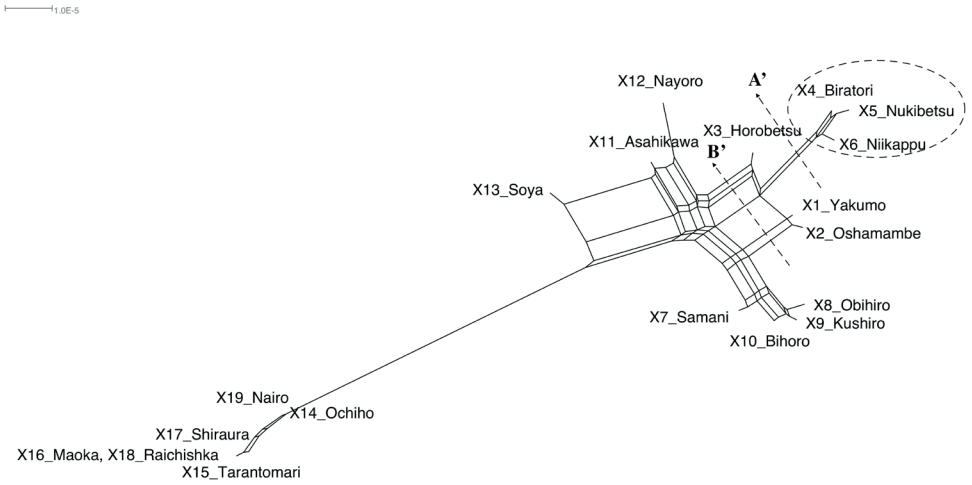

Figure 3. The result of Neighbor-Net Analysis applied to the Euclidean distance matrix for Biratori dialect of Homogeneity Analysis assuming the nominal scale in Hattori and Chiri (1960).

which correspond to the Euclidean coordinates in Figure 7.

Furthermore, the result of cluster analysis on the left in Figure 9 (i.e., on the ordinal scale) demonstrates the unique position of Samani dialect as a dendrogram in contrast to the northeastern grouping on the right in Figure 9 (i.e., on the nominal scale).

These findings, which were obtained from statistical analysis with the assumption of the ordinal scale, are consistent with the current linguistic and philological knowledge presented in Section 2.3 and suggest that previous statistical analysis with the assumption of the nominal scale has obscured the underlying information

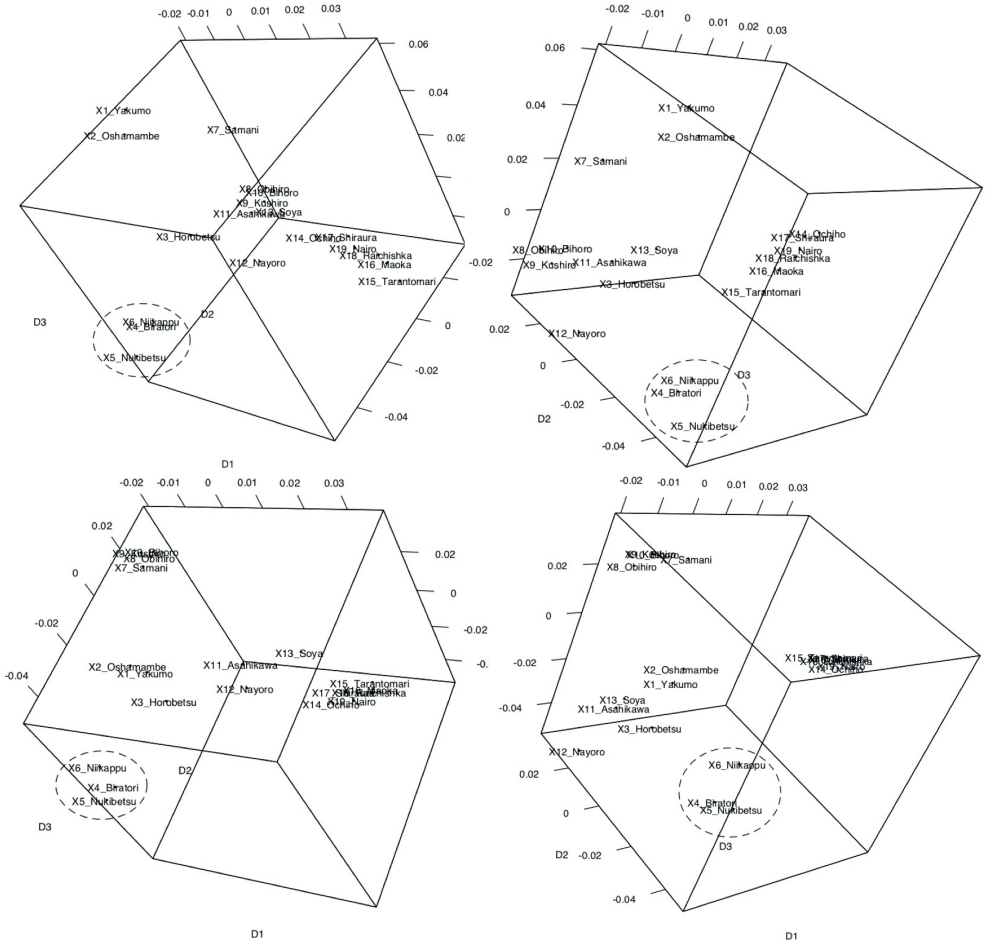

Figure 4. The result of three-dimensional plots for the Euclidean coordinate system for Biratori dialect yielded by Homogeneity Analysis in Hattori and Chiri (1960).
(Top) on ordinal scale (Bottom) on nominal scale

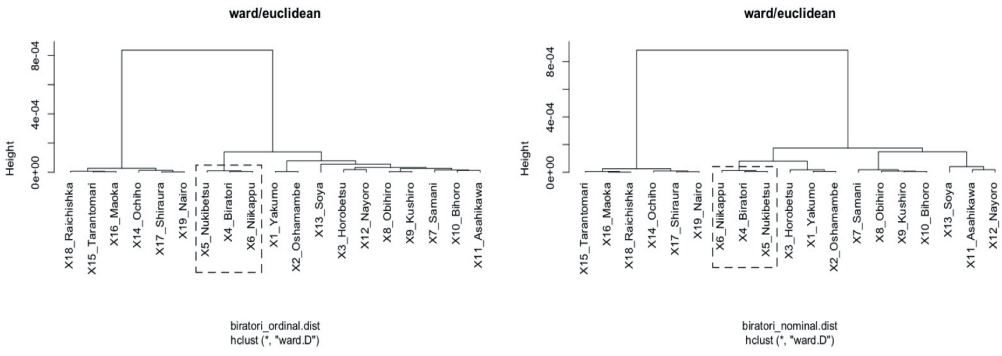

Figure 5. The results of cluster analysis of Euclidean distance for Biratori dialect.
(Left) on ordinal scale (Right) on nominal scale.

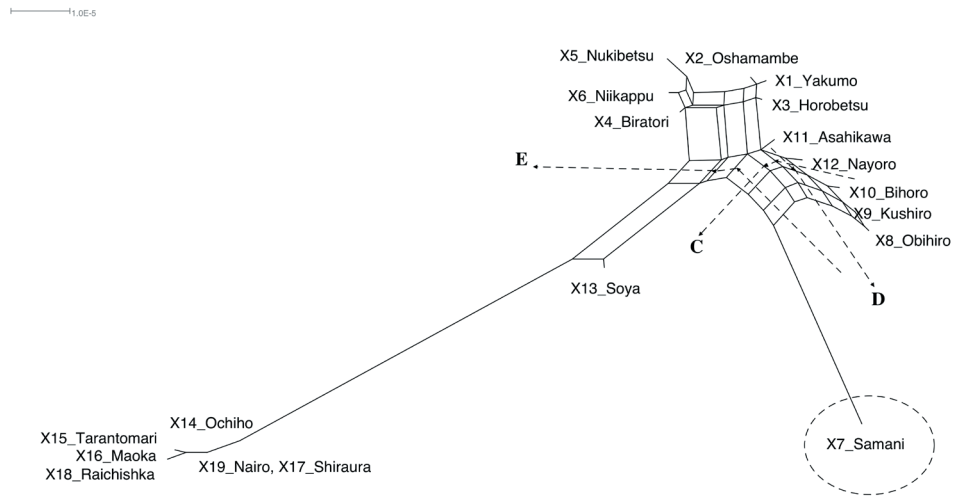

Figure 6. The result of Neighbor-Net Analysis applied to the Euclidean distance matrix for Samani dialect of Homogeneity Analysis assuming the ordinal scale in Hattori and Chiri (1960).

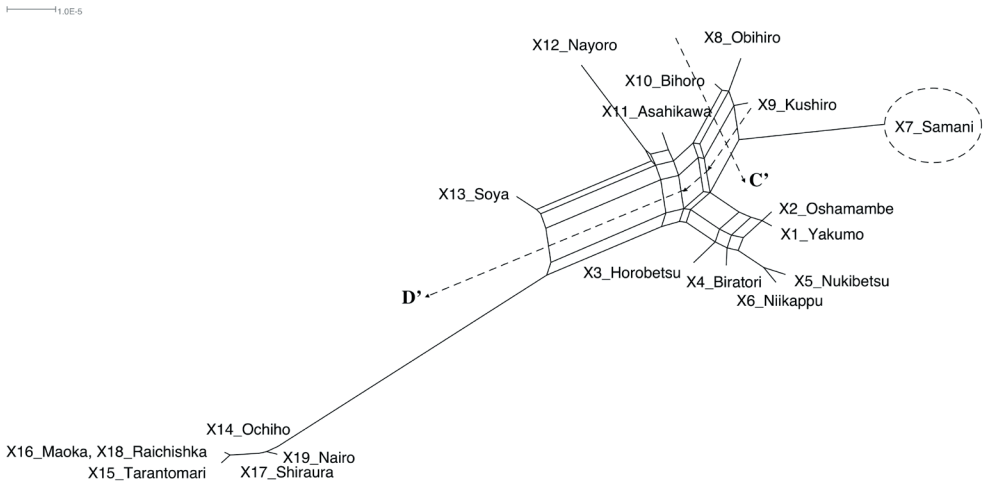

Figure 7. The result of Neighbor-Net Analysis applied to the Euclidean distance matrix for Samani dialect of Homogeneity Analysis assuming the nominal scale in Hattori and Chiri (1960).

structures (i.e., dialect relationships in this case), which the five symbols (i.e., "±," "◯," "?," "・," and "()") have borne in the half century since Hattori and Chiri (1960) have left the reality of the investigation of Ainu dialects in those days to these symbols.

4. Discussions and Conclusion

In this Section, the author discusses what the main results in this paper suggest for current and future linguistic research (including lexicostatistics) from the viewpoints of both statistics and linguistics.

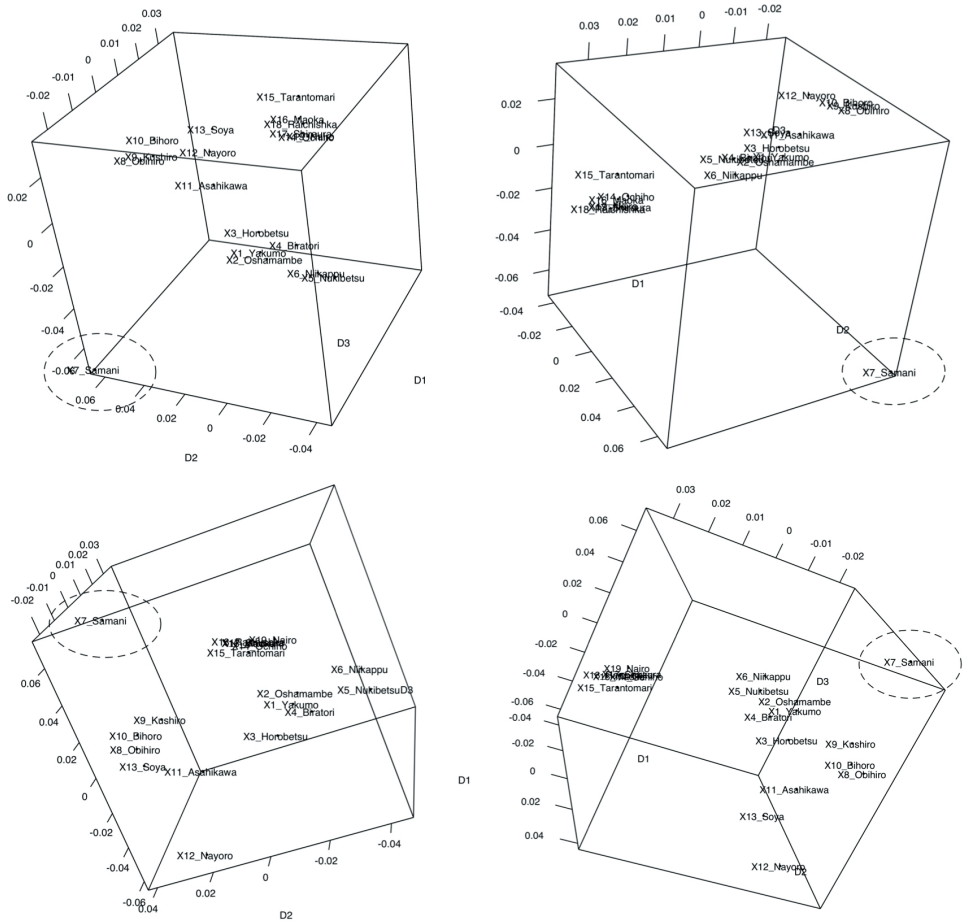

Figure 8. The result of three-dimensional plots for the Euclidean coordinate system for Samani dialect yielded by Homogeneity Analysis in Hattori and Chiri (1960).
(Top) on ordinal scale (Bottom) on nominal scale

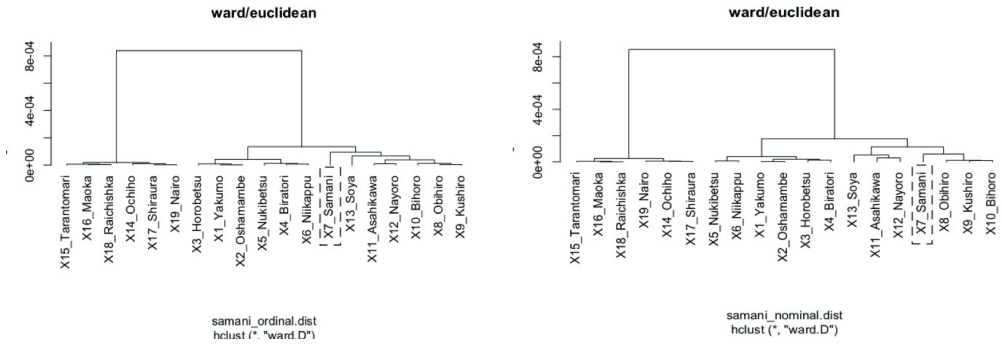

Figure 9. The results of cluster analysis of Euclidean distance for Samani dialect.
(Left) on ordinal scale (Right) on nominal scale

From the statistical point of view, the results of this paper indicate that lexicostatistical surveys, which cannot necessarily avoid or compensate for "some uncertainty" in the original records (e.g., the linguistic environment around Ainu language in the 1960's in our case), may contain some significant information in "some uncertainty" itself and that the recent computational and statistical advancements have already enabled researchers to quantify lexicostatistical data that includes "some uncertainty" more appropriately than before. Therefore, the author recommends that researchers apply statistical analysis with the assumption of the ordinal scale (e.g., Homogeneity Analysis with the ordinal scale) to data with "some unavoidable uncertainty."

Furthermore, the statistical reanalysis of the previous lexicostatistical research for an endangered language like Ainu could bring new insights to the language by utilizing appropriate statistical analysis (e.g., Homogeneity Analysis with ordinal assumption in this paper) to the present data. Moreover, the scope of the attempt could be beyond the lexicostatistics. Again, in Section 2.2, linguistic data, in general, are recorded with many "symbols." Linguistic research based on these symbols then requires some quantification methods (i.e., to assign appropriate values to these symbols) in practice. Thus, many linguistic researchers are confronted with this sort of quantification problem with some uncertainty in the data, consciously or implicitly. Therefore, complementary studies among linguists and statisticians will be a promising approach in linguistics.

From the linguistic point of view, the main results of this paper are summarized in the following three points. First, from the substantive knowledge on Ainu dialectology, the result is noted that the data based on cognacy judgments for a given dialect (i.e., Biratori or Samani in this paper) and other dialects make the dialect classification more consistent with Ainu dialectology.

Since Asai (1974) classified Hokkaido Ainu dialects into four parts: Central South Hokkaido group typified by Nos. 1-6 in Figure 1, Eastern Hokkaido group typified by No. 7, East Hokkaido group typified by Nos. 8-12, and North Hokkaido group typified by No. 13, his classification has great influence on Ainu linguistics.

Notably, Asai (1974) demonstrated that (1) Hokkaido Ainu dialects are clustered into southwestern Hokkaido Ainu group (i.e., Nos. 1-6 in Figure 1) and northeastern Hokkaido Ainu group (i.e., Nos. 7-13 in Figure 1), and that (2) Samani dialect is classified into northeastern Hokkaido Ainu group. The first point has been inconsistent with Saru-Chitose-type (Nakagawa 1996: 11) because of the geographical locations similar to Eastern and Western patterns (i.e., northeastern and southwestern classification in our case) and Center versus Periphery patterns (i.e., Saru-Chitose-type in our case) in Japanese dialectology. The second point also has contradicted the philological research that indicated the importance and uniqueness in and around Samani dialect.

However, the main results in this paper illustrate that the previous statistical analyses with the assumption of the nominal scale in Hattori and Chiri (1960) have obscured the underlying relationships of these two dialects and the statistical analyses with the assumption of the ordinal scale have demonstrated (1) Saru-Chitose-type from the viewpoint of Biratori dialect and (2) the distinct characteristics of Samani dialect from the viewpoint of Samani dialect.

These results suggest that lexicostatistical data in Hattori and Chiri (1960) contain significant information on the ordinal scale and the statistical analysis with the assumption of the ordinal scale (i.e., Homogeneity

Analysis with ordinal assumption in this paper) succeeds in capturing the dialect relationships.

Therefore, it is recommended that linguistic researchers record not only what is linguistically distinguishable, but also, according to the circumstances, some "unavoidable" uncertainty in the investigation as it is. As the main results in this paper demonstrate, those records have the potential to shed new light on the linguistic research[9].

Second, the results in this paper demonstrate that the linguistic research of dialect or language with "some (serious and) unavoidable uncertainty" in the data necessarily have and need to pay more attention to "the problem" that the classification of dialects or languages could be affected by which dialect or language to focus on or to study.

In fact, the overall relationships of the 19 Ainu dialects (e.g., the result of Neighbor-Net) from the viewpoints of Biratori dialect are different from the viewpoints of Samani dialect in some respects. As discussed in Section 1, "some uncertainty" regarding the lexicostatistical data in Hattori and Chiri (1960), as represented by the five symbols (i.e., "±," " ○ ," "?," " • ," and "()"), has caused these differences in contrast to the "normal" lexicostatistical data like Table 1[10].

Third, the main results in this paper may pose questions on the current application of statistical methods to phylogenetic research including Bayesian Phylogenetic Analysis. Since the linguistic research on language phylogeny applies the phylogenetic methods developed in evolutionary biology with an assumption of DNA (i.e., a set of discretized elements [A, G, C, and T or A, G, C, and U]), the simple application of these methods to linguistic data can lead to the violation of basic assumptions in biological data that causes unnecessary confusion in linguistics.

As Homogeneity Analysis assigns continuous values to the data with ordinal assumption, the main results in this paper that the application of Homogeneity Analysis succeeds in visualizing the underlying information structures in Hattori and Chiri (1960) suggest a reconsideration of the present linguistic research on language phylogeny utilizing the phylogenetic methods in evolutionary biology, and a need of new statistical methodologies that can apply to the data with continuous values (e.g., how to implement the mutation system in the data with continuous values).

Finally, the author discusses the background of this paper: why the statistical analysis assumes from the historical background that the ordinal scale in the data has not been widely used in either the humanities or sciences. To conclude, the author will then remark on the future directions of complementary studies among

9) One of reviewers suggested a possibility that the proposed methods could apply to the whole humanities. Since humanities research is, in general, confronted with "some unavoidable uncertainty" in various circumstances, the results in this paper that researchers can utilize "some unavoidable uncertainty" itself for the matter of concern indicate that interdisciplinary research among the humanities and statistics will bring new insights into the humanities.

10) Sato (2008: 153-156) has also pointed out the same problem in the context of the research for old documentation in Ainu. If the author, as a statistician, dares to transfer his insightful view into a statistical context, his issues are closely related to whether a statistical approach to linguistic data with "some (serious and) unavoidable uncertainty" enables us to integrate the whole relationships among dialects or languages from the viewpoint of each dialect or language as "one whole relationship" and, if possible, what statistical analysis is most appropriate for the analysis. Due to space limitations, the author will deal with this interesting and novel issue in another article.

the humanities and statistics.

The origin of the "quantification method" may go back to both psychology and sciences. The realm of psychology has today developed as "mathematical psychology" or "psychometrics." Psychologists are first confronted with how to measure or quantify what "psychological" is and then what operation (e.g., addition, subtraction, multiplication, division, order relation, or counting) can be applied to the measured or quantified "psychological." The main problem is in the different properties of what is "psychological."

For example, if a psychologist, whose research question is the attitude of the respondent to some political issues, uses a questionnaire in his research, and the respondent records his attitude to a political issue on a five-point (Likert) scale (i.e., [1]: strongly agree, [2]: agree, [3]: neutral, [4]: disagree, and [5]: strongly disagree), one can easily have an idea (even some [psychological] researchers) that the five-level (Likert) items are quantified in some manner (e.g., assigning 2, 1, 0, -1, and -2 to [1]: strongly agree, [2]: agree, [3]: neutral, [4]: disagree, and [5]: strongly disagree) and calculate some statistics (e.g., mean, deviation, or variance) based on the scores. In this case, researchers are supposed to assume the strong relations of the "psychological" attitude to a political issue: addition, subtraction, multiplication, division, order relation, and counting.

However, real data often violate even order relation. For example, answer (1): strongly agree and answer (5): strongly disagree seem to be the farthest among the five-level (Likert) items above but these answers are similar in the way that the political issue is of concern for the respondent; that is, not indifferent. In practice, many psychologists often suffer from the nonlinear effect (i.e., answer [3]: neutral and the two answers above are much farther in terms of "indifference") in the results of such questionnaire data. Therefore, the main focus of the "quantification method" in psychology has, to some degree, been on such a phenomenon violating even the order relation: the data on the nominal scale. Furthermore, the recent application of quantification methods to linguistic data has revealed that the nonlinear effect occurs even in linguistic typology (Ono, Yoshino, Hayashi, and Whitman 2017; Ono, Yoshino, Hayashi, and Whitman 2018). Thus, it is natural that quantification methods in psychology have developed with the assumption of the nominal scale in the data.

However, the recent extension of statistical analyses to the humanities has opened up a new and unknown field for statistics, as the main results in this paper indicate. Humanities data, including the process of how humanities researchers record and quantify the data consciously or implicitly, contain a complex mechanism in ways that statistical analysis developed in psychology or science does not presuppose. Thus, statisticians addressing humanities data should reconsider our assumption present in the statistical analysis and develop new statistical methodologies, which might relax the current statistical assumption, add some new aspects to humanities research, and bring new insights into the humanities.

Therefore, the author has written this paper as a starting point of this new interdisciplinary research, taking as an example the ordinal scale in lexicostatistics in Hattori and Chiri (1960). The author wishes to end this paper with the hope that the complementary studies among the humanities and statistics will prove a fruitful discipline in the future.

Acknowledgments

The author is grateful to the editors and two highly conscientious reviewers; all errors are of course my own.

References

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

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

de Leeuw, J., and Mair, P. (2009) Gifi method for optimal scaling in R: The package homals. *Journal of Statistical Software, 31*(4), 1-20.

Geospatial Information Authority of Japan. (2018) Ministry of Land, Infrastructure, Transport and Tourism. https://maps.gsi.go.jp [accessed on November 2018].

Gifi, A. (1990) *Nonlinear multivariate analysis*. John Wiley and Sons.

Hattori, S., and Chiri, M. (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. (in Japanese)

Hayashi, C. (1952) On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 3*(1), 69-98.

Huson, D., and Bryant, D. (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution, 23*(1), 254-267. Oxford: Oxford University Press.

Inoue, F. (1985/8) *Atarashii nihongo: shinhougen no bunpu to henka [The Changing Japanese Language: Distribution and Change of New Dialect Forms]*. Tokyo: Meiji Shoin. (in Japanese)

Lee, S., and Hasegawa, T. (2013) Evolution of the Ainu language in Space and Time. *PloS One, 8*(4), e62243.

Likert, R. (1932) A Technique for the Measurement of Attitudes. *Archives of Psychology, 22 140*, 55.

Nakagawa, H. (1996) Gengo chirigaku ni yoru ainugo no shiteki kenkyū [The historical study of Ainu from the viewpoint of geolinguistics]. *Bulletin of the Hokkaido Ainu Culture Research Center, 2*, 1-17. (in Japanese)

Nishisato, S. (2006) *Multidimensional nonlinear descriptive analysis*. Chapman and Hall/CRC.

Ono, Y. (2015) Statistical Reanalysis about the Classification of Ainu Dialects: On the Data of Hattori and Chiri (1960), *Journal of the Center for Northern Humanities, 8*, 25-41. (in Japanese)

Ono, Y. (2019) Some Problems on the Lexicostatistical Data and the Extension of Lexicostatistics in New Directions. Manuscript in preparation.

Ono, Y., Yoshino, R., Hayashi, F., and Whitman, J. (2017) A Multiple Correspondence Analysis of the Latent Structure of Features in Linguistic Typology (1): A Statistical Reanalysis of Tsunoda, Ueda, and Itoh (1995a), *Mathematical Linguistics, 31*(3), 189-204.

Ono, Y., Yoshino, R., Hayashi, F., and Whitman, J. (2018) A Multiple Correspondence Analysis of the Latent Structure of Features in Linguistic Typology (2): A Statistical Reanalysis of Tsunoda, Ueda, and Itoh (1995a), *Mathematical Linguistics, 31*(4), 261-280.

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical

Computing, Vienna, Austria. https://www.R-project.org/.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1983) *Jōhōryō toukeigaku [Statistics on Information Criteria]*, Tokyo: Kyōritsu Shuppan. (in Japanese)

Sato, T. (2002) A Basic Vocabulary of the Samani Dialect of Ainu. *The Annual Report on Cultural Science, 106*, 91-126.

Sato, T. (2008) Ainugo kobunken ni okeru gengogakuteki shomondai [Some linguistic problems on the research of old documentation in Ainu], *The Annual Report on Cultural Science, 124*, 153-180. (in Japanese)

Refsing, K. (1986) *The Ainu language: the morphology and syntax of the Shizunai dialect*. Aarhus: Aarhus University Press.

Swadesh, M. (1955) Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics, 24*, 121-137.

Tamura, S. (1970) Personal affixes in the Saru dialect of Ainu, in Jakobson, Roman and S. Kawamoto (eds.). *Studies in general and oriental linguistics*, presented to Hattori Shirō, 577-611. Tokyo: TEC Company Ltd.

Tamura, S. (2000) *The Ainu Language* (ICHEL Linguistic Studies 29) Tokyo: Sanseidō.

Abstract

Linguistic data are generally recorded using many “symbols.” Therefore, linguistic research based on these symbols is, in practice, confronted with a sort of quantification problem (i.e., to assign appropriate values to these symbols), consciously or implicitly. However, humanities data, including the process of how a researcher in the substantive field records the matter of concern and quantifies those records, contain a complex mechanism in various ways that current statistical methods developed in psychology and science do not necessarily assume. Hence, statisticians addressing humanities data should reconsider the assumptions present in statistical analysis and develop new statistical methodologies, which will relax the present assumption, add some new aspects to the humanities, and bring new insights into the humanities.

As its starting point, this paper is an attempt to address this new interdisciplinary research among the humanities and statistics, taking lexicostatistical data in Hattori and Chiri (1960) as an example. The five symbols (i.e., “±,” “ ○,” “?,” “ • ,” and “()”), which Hattori and Chiri (1960) introduced to record some unavoidable uncertainty in the linguistic environment of Ainu in the 1960’s, led the author to extend the data type for lexicostatistics and the hypothesis of the ordinal scale in lexicostatistics. The application of statistical analysis with the assumption of the ordinal scale clearly demonstrates the significance and uniqueness in and around Biratori and Samani dialects.