Northern Language Studies 15: 179-202 (Japan Association of NoLS, 2025)

Statistical Approaches to the Quantification of Regularity in Languages and Dialects:

An Exercise in Ainu Dialects of Asai (1974)

Yohei ONO
(St. Luke's International University)
Mika FUKAZAWA
(National Ainu Museum)

Keywords: Ainu language, phonological correspondence, feature engineering, multidimensional visualization, quantitative descriptive linguistics

1. Introduction

The information of "regularity," so-called phonological correspondences, morphological and grammatical rules, is inherently present in the lexical data of languages and dialects[1]. However, it cannot always be analyzed in a numerical way that is applied to lexical items and forms[2], as seen in previous linguistic research (e.g., lexicostatistics). Previous studies (Ono and Fukazawa 2023, 2024) have shown that the basic vocabulary containing phonological correspondences in Ainu dialects were excluded from Asai's (1974) classification.

Herein, our main objective is not to obtain the "correct" classification in the Ainu dialects, which can vary based on the classification purpose and the corresponding criteria in cognacy judgments, as shown in Ono and Fukazawa (2024). Rather, this paper explores the possibilities for researchers to extract and quantify regularities in languages and dialects using Asai's (1974) data on the Ainu dialects.

Accordingly, we aim to evaluate whether the extracted and quantified regularities, previously excluded by Asai (1974) for being the same lexical form with phonological correspondences rather than different lexical forms, can not only duly increase the information that cannot be captured by the lexicostatistical framework but also further clarify the characteristics of Asai's (1974) classification of the Ainu dialects.

The remainder of this paper is organized as follows: Section 2 introduces the

[1] This paper will introduce terminologies of "regularity" and "type," as "regularity" is comprised of several "types." For example, the semivowel (glide) regularity is comprised of *iw*-type and *uy*-type extracted from two basic items of the Ainu language, 179.pierce (stab): *ciw*, *cuy*; 74.star: *nociw*, *noociw*, *nocuy* in Asai (1974), as discussed in Section 3.

[2] For example, lexical item corresponds to "star," and lexical form corresponds to *nociw*.

background of Asai's (1974) study and classification of the Ainu dialects, coupled with some examples of basic items containing regularity thereof, resulting in their exclusion from Asai's (1974) analysis.

Section 3 addresses the issues of extracting regularity information from the Ainu dialect materials from the viewpoints of descriptive linguistics and assigning numbers to the extracted regularity, while considering the possibility of dealing with the uncertainty of the linguistic materials as appropriately as possible. Accordingly, we demonstrate that our methods are capable of extracting a greater degree of information pertaining to regularity. Actually, lexicostatistics cannot extract this information owing to its limitation of needing to separate and classify the lexical forms in the item only in a single, but not multiple, manner. Several uncertainties are also discussed in the context of the research on the linguistic materials of the Ainu dialects.

Section 4 focuses on how to statistically analyze the extracted regularity, whereby the statistical methods aim to enable researchers to preserve the context of linguistic materials regarding quantified regularity. First, correspondence analysis (Benzécri et coll. 1973), which is generally used to quantify lexical forms in geolinguistics, is considered as inappropriate for quantifying the extracted regularity. Second, we propose homogeneity analysis (Gifi 1990) as an alternative method. Third, the linguistic requirements discussed in Section 3 lead us to revise the present algorithm of homogeneity analysis in R language (de Leeuw and Mair 2009), and our modifications are explained.

Section 5 applies our revised homogeneity analysis to the data of Asai's (1974) 110 items with and without the extracted regularity, and compares the classification results of the Ainu dialects from linguistic and statistical perspectives. Consequently, we confirm that the extracted and quantified regularities enhance the information that cannot be captured by the lexicostatistical framework and contribute to clarifying Asai's (1974) original classification of the Ainu dialects. These findings have led to the clarification of the characteristics of Asai's (1974) Ainu dialect classification and suggested that the regularity information, which is analyzed from Asai's (1974) perspective, can be appropriately extracted and quantified via our proposed methods.

Section 6 discusses the significance of our results from a linguistic and statistical standpoint.

2. Background

As explained in Ono (2020: 37), two studies exist on basic vocabulary in Ainu dialects: one by Hattori and Chiri (1960) and another by Asai (1974). Hattori and Chiri (1960) conducted lexicostatistical surveys of 19 Ainu dialects (marked by Nos. 1–19 in Figure 1) and collected lexical information on 200 basic vocabulary items in each dialect with reference to Swadesh's basic vocabulary list (cf., Fukazawa and Ono 2025). Since most Ainu dialects were on the verge of vanishing at that time, their research made an

invaluable contribution to Ainu linguistics.

Furthermore, Asai (1974) has added some revisions to the lexicostatistical data in Hattori and Chiri's (1960) study on the Obihiro, Kushiro, and Asahikawa dialects (Nos. 8, 9, and 11 in Figure 1) based on his own fieldwork (Asai 1974: 66); collected the data of Chitose dialect (No. 21 in Figure 1) from informants (Asai 1974: 64–66); and gathered the North Kuril dialectal data (No. 20 in Figure 1) with reference to Torii (1903), Murayama (1971), and Pinart's vocabulary manuscripts (Asai 1974: Appendix). Moreover, Asai (1974) corrected his data based on Hattori (1964) as needed.

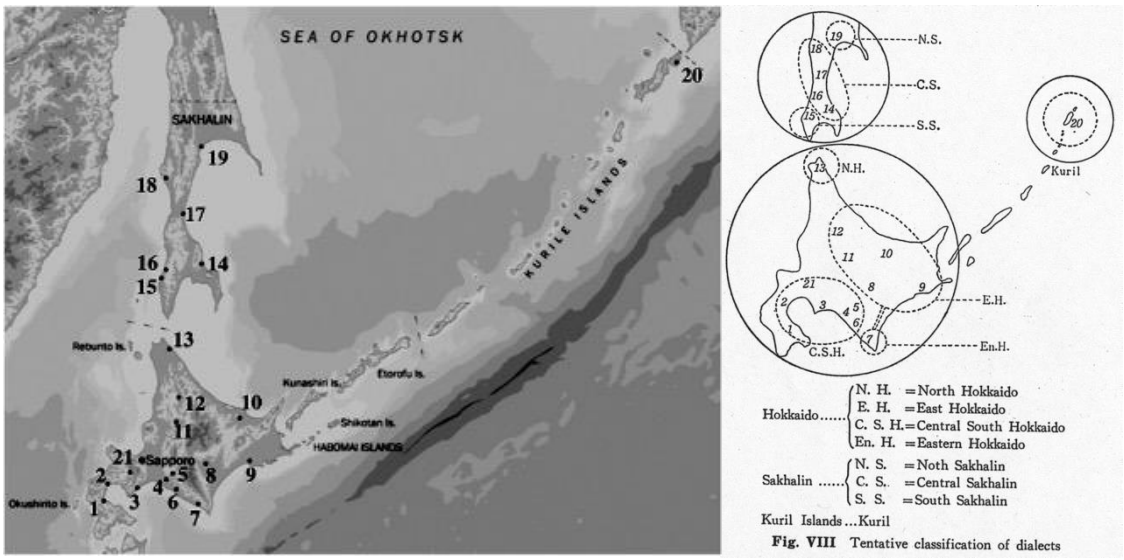


Figure 1: Left: Map of the Ainu dialects (Geospatial Information Authority of Japan, 2019). Note: 1: Yakumo, 2: Oshamambe, 3: Horobetsu, 4: Biratori, 5: Nukibetsu, 6: Niikappu, 7: Samani, 8: Obihiro, 9: Kushiro, 10: Bihoro, 11: Asahikawa, 12: Nayoro, 13: Soya, 14: Ochiho, 15: Tarantomari, 16: Maoka, 17: Shiraura, 18: Raichishka, 19: Nairo, 20: North Kuril (Shumushu), 21: Chitose. Right: Asai's classification of the Ainu dialects (Asai 1974: 100).

Hattori and Chiri (1960) and Asai (1974) have attempted to classify the Ainu dialects, applying some statistical methods to the data of 200 (or Asai's 202) basic vocabulary items. Consequently, Asai's (1974) classification, shown in the map on the right side of Figure 1, has been cited as a landmark in the studies of the Ainu language.

However, previous studies (Ono 2020; Ono and Fukazawa 2023, 2024) have demonstrated that the two classifications by Hattori and Chiri (1960) and Asai (1974) have been conducted based on the different lexical items and cognacy judgments among the lexical forms, which had been unidentified, making it impossible for researchers to compare and examine both classifications from the viewpoint of contemporary studies of the Ainu dialects.

Furthermore, of the 202 basic vocabulary items, only 110 items were used for, while 92 basic vocabulary items of the Ainu dialects were excluded from Asai's (1974) classification partly due to his cognacy judgments on lexical forms. For example, as in Ono and Fukazawa (2024) and Fukazawa and Ono (2025), the items, such as 54.drink: (*ku* (1–13, 20, 21), *kuu* (14–19)) and 83.ashes: *uyna* (1–3), (*una* (4–13, 20, 21), *uuna* (14–19)), were excluded from the 110 items in Asai (1974) because they were considered as cognates depending on the phonological regularities between lexical forms. The main objective of this paper is to provide some possibilities for researchers to extract and quantify the regularities in languages and dialects using Asai's (1974) data of the Ainu dialects.

Thus, the following section will validate our proposed methods for extracting and quantifying the regularities. We particularly assess whether the extracted and quantified regularities can enhance information that cannot be captured by the lexicostatistical framework. Furthermore, we investigate whether the extracted and quantified regularities, excluded from Asai (1974) on the grounds that they are "regularity" rather than "lexical form" but reflected in his analysis of phonological correspondence in the Ainu dialects as discussed in this section, can provide greater clarity regarding the characteristics of Asai's (1974) classification of the Ainu dialects.

3. Data: How to Extract "Regularity"

This section introduces the regularity information in Asai (1974) that is mainly dealt with in Fukazawa and Ono (2025) and addresses how to numerically extract the regularity, with a particular focus on the challenges encountered in the process behind Fukazawa and Ono (2025).

In principle, let us summarize the terminologies related to regularity in the following section. Table 1 shows the semivowel (glide) regularity extracted from two basic items in Asai (1974): 179.pierce (stab): *ciw* (3, 4, 8, 9, 12–17, 19), *cuy* (10); 74.star: (*nociw* (1–9, 11, 12, 21), *noociw* (19), *nocuy* (9))[3]. The following part of this paper assumes "regularity" to be comprised of several "types." For example, the semivowel (glide) regularity consists of the *iw*- and *uy*-types in Table 1. As explained in Fukazawa and Ono (2025), this regularity can be explained as a metathesis between the *iw*- and *uy*-types.

There are novel issues with respect to the extraction of regularity and the corresponding types in this paper and Fukazawa and Ono (2025). In our previous studies (Ono and Fukazawa 2023, 2024) on lexicostatistical data in Asai's (1974) Ainu dialects, lexicostatistics is primarily concerned with the classification of lexical forms that are applicable to a single item but not multiple items, which is the basic assumption of the

[3] The number in parentheses after lexical form corresponds to the place in Figure 1, and the lexical forms in the same parentheses correspond to those that Asai (1974) judged to be cognates, as recovered by our previous studies (Ono 2020; Ono and Fukazawa 2023, 2024).

Table 1: The semivowel (glide) regularity.

	iw-type	*uy*-type
1_Yakumo	1	0
2_Oshamambe	1	0
3_Horobetsu	1	0
4_Biratori	1	0
5_Nukibetsu	1	0
6_Niikappu	1	0
7_Samani	1	0
8_Obihiro	1	0
9_Kushiro	1	1
10_Bihoro	0	1
11_Asahikawa	1	0
12_Nayoro	1	0
13_Soya	1	0
14_Ochiho	1	0
15_Tarantomari	1	0
16_Maoka	1	0
17_Shiraura	1	0
18_Raichishka	0.5	0.5
19_Nairo	1	0
20_North Kuril	0.5	0.5
21_Chitose	1	0

110 items in the Ainu dialects of Asai (1974) used in our analysis. However, the regularity is typically extracted using multiple items.

For example, the semivowel (glide) regularity in Table 1 is extracted from two items in Asai (1974): 179.pierce (stab): *ciw* (3, 4, 8, 9, 12–17, 19), *cuy* (10); 74.star: (*nociw* (1–9, 11, 12, 21), *noociw* (19), *nocuy* (9)).

The three requirements are as follows. First, the process of establishing regularity corresponds to organizing the presence of types in the regularity as the value of 1. Second, the presence of types will increase during the process of establishing regularity; however, the absence of types, represented by the value of 0, will not. Third, the quantification of the extracted regularity should focus on the presence of types rather than their absence. The absence of types indicates that the presence of types cannot be observed in the dialect. Consequently, the absence of types cannot be generally demonstrated from a linguistic perspective.

These three requirements are called 1) the principle of organizing the presence; 2) the principle of increasing the presence; 3) the principle of quantifying the presence. In the following subsections, we will explain these three requirements using the example of semivowel (glide) regularity, and determine if the extracted regularity can actually enhance information in the Ainu dialects not captured by lexicostatistics.

3.1 The First Requirement: The Principle of Organizing the Presence

The main objective of this subsection is to explain the first requirement in extracting regularity, that is, the principle of organizing the presence, taking the semivowel (glide) regularity as an example.

Table 2 shows the tentative semivowel (glide) regularity just extracted from the item of 179.pierce (stab): *ciw* (3, 4, 8, 9, 12–17, 19), *cuy* (10). In extracting the semivowel (glide) regularity using the first item, we confirm the *iw*-type in the Horobetsu, Biratori, Obihiro, Kushiro, Nayoro, Soya, Ochiho, Tarantomari, Maoka, Shiraura, and Nairo dialects as *ciw*, and the *uy*-type in the Bihoro dialect as *cuy*, resulting in setting the corresponding values as 1.

In establishing semivowel (glide) regularity, it is possible to extract only the presence of the corresponding type (e.g., a value of 1 in the *iw*- and *uy*-types) in the linguistic

Table 2: The tentative semivowel (glide) regularity.

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

Table 3: The updated semivowel (glide) regularity.

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

materials of each dialect. However, it is impossible to extract the absence of the corresponding type (e.g., a value of 0 in the *iw*- and *uy*-types).

Thus, the first requirement is referred to as the principle of organizing the presence in this paper. Regarding this principle, Table 2 shows that the *iw*-type exists in the linguistic material on the Kushiro dialect, but the table does not show that there is no *iw*- or *uy*-type in the linguistic material on the Yakumo dialect. Thus, the value of 0 in Table 2 corresponds to the result that we cannot confirm any *iw*- or *uy*-type from the linguistic material (i.e., item of 179.pierce (stab)).

3.2 The Second Requirement: The Principle of Increasing the Presence

The main objective of this subsection is to explain the second requirement in extracting regularity, that is, the principle of increasing the presence. As explained in the previous subsection, Table 2 shows the tentative semivowel (glide) regularity extracted from the first item of 179.pierce (stab). Table 3 shows the semivowel (glide) regularity updated by the second item of 74.star: (*nociw* (1–9, 11, 12, 21), *noociw* (19), *nocuy* (9)).

In updating the semivowel (glide) regularity by the second item, we confirm the *iw*-type in the Yakumo, Oshamambe, Nukibetsu, Niikappu, Samani, Asahikawa, and Chitose dialects as *nociw*, and the *uy*-type in the Kushiro dialect as *nocuy*, resulting in the revision of the corresponding values to 1.

In updating Table 2, it is only possible to increase the presence of the corresponding type (e.g., a value of 1 in the *iw*- and *uy*-types) in the linguistic materials of each dialect as the gray cells in Table 3. However, it is impossible to increase the absence of the corresponding type (a value of 0 in the *iw*- and *uy*-types).

Thus, in this study, the second requirement is called as the principle of increasing the presence, which is based on the fact that we can only observe and update the presence of the corresponding type in the linguistic materials of each dialect but cannot "demonstrate" the absence of the corresponding type in most cases, as explained in the next subsection.

3.3 The Third Requirement: The Principle of Quantifying the Presence

The main objective of this subsection is to explain the third requirement in extracting regularity, that is, the principle of quantifying the presence. Table 4 shows the final semivowel (glide) regularity extracted from the two lexical items of 179.pierce (stab) and 74.star.

As explained in Sections 3.1 and 3.2, the value of 0 does not correspond to demonstrating no *iw*- or *uy*-type in the linguistic material of the dialect, but the value of 0 corresponds to the fact where we cannot observe or determine the corresponding type in the linguistic material of the dialect. For example, the gray cell in Table 4 does not show that there is no *uy*-type in the linguistic material of the Obihiro dialect but that we cannot observe the *uy*-type in the final semivowel (glide) regularity using the two lexical items.

Table 4: The final semivowel (glide) regularity.

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

Therefore, the statistical methods applied to the extracted regularity and the corresponding type are required for quantifying only the presence of types (i.e., the value of 1) but not the absence of types (i.e., the value of 0) in Table 4.

Thus, the third requirement is termed the principle of quantifying the presence. The principle of quantifying the presence is based on the fact that we can only observe the presence of the corresponding type in the linguistic materials of each dialect and the reliable information is only the presence of the type in the regularity but not the absence of the type in most cases, as explained in the previous subsection[4].

[4] There remain issues regarding how to deal with the information that there are no types in the regularity observed in the linguistic materials of a particular dialect. For example, there are no *iw*- or *uy*-type in the Raichishka and North Kuril dialects in Table 4 because we cannot observe the corresponding form related to the *iw*- or *uy*-type in the basic vocabulary of Asai (1974). Herein, we tentatively replaced the value of 0 in the dialect containing no types in the regularity by 1 over the number of types in the corresponding regularity. For example, the value of 0 was replaced by 1 over the number of types in the semivowel (glide) type (i.e., 1 over 2) in the Raichishka and North Kuril

3.4 How Extracted Regularity Can Increase the Information Not Captured by Lexicostatistics

In the preceding subsections, three requirements were explained for extracting regularity. The main objective of this subsection is to demonstrate that the extracted regularity can increase the degree of information not captured by lexicostatistics, taking the lexical forms of 19.fish and the regularity of the V_1V_2 and V_2 types extracted from the item of 19.fish as an example.

Let us focus on the example of the corresponding lexical forms of 19.fish in Asai (1974) as in Table 5. There are three lexical forms, that is, *cep*, *ciep*, and *ceh,* in the item of 19.fish in Asai (1974).

Table 5: The corresponding lexical forms on 19.fish in Asai (1974).

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

Table 6: Asai's cognacy judgments on 19.fish.

	(*cep*, *ciep*)	*ceh*
1_Yakumo	1	0
2_Oshamambe	1	0
3_Horobetsu	1	0
4_Biratori	1	0
5_Nukibetsu	1	0
6_Niikappu	1	0
7_Samani	1	0
8_Obihiro	1	0
9_Kushiro	1	0
10_Bihoro	1	0
11_Asahikawa	1	0
12_Nayoro	1	0
13_Soya	1	0
14_Ochiho	0	1
15_Tarantomari	1	0
16_Maoka	0	1
17_Shiraura	0	1
18_Raichishka	0	1
19_Nairo	1	0
20_North Kuril	1	0
21_Chitose	1	0

As previous studies (Ono 2020, Ono and Fukazawa 2023, 2024) have demonstrated in Table 6, Asai (1974) classifies these three lexical forms into (*cep*, *ciep*) and *ceh*, which distinguish the phonological correspondence between *-p* and *-h* as non-cognate but that between *-e-* and *-ie-* as cognate.

The lexicostatistical viewpoint of Asai (1974) has been limited in the point that

dialects in Table 4, which could be considered as the dialect containing the semivowel (glide) regularity from the viewpoint of linguistics, but either the *iw-* or *uy*-type was not observed in the linguistic materials of the dialects used in this paper, that is, the basic vocabulary of Asai (1974). Further investigations are needed regarding this issue in future research.

lexicostatistics needs to separate and classify lexical forms in an item only in a single, but not multiple, manner. For example, in Table 5, researchers must select one of five cognacy judgments for the three lexical forms: *cep*/*ciep*/*ceh*, (*cep*, *ciep*)/*ceh*, *cep*/(*ciep*, *ceh*), *ciep*/(*cep*, *ceh*), and (*cep*, *ciep*, *ceh*), the last of which corresponds to the cognacy judgment in Hattori and Chiri (1960). In other words, researchers cannot adopt the multiple cognacy judgments for the three lexical forms in the lexicostatistics. For example, we cannot choose either (*cep*, *ciep*)/*ceh* or (*cep*, *ceh*)/*ciep* in the lexicostatistics simultaneously, which is a fundamental limitation of the information extracted in the lexicostatistical framework.

However, the extracted regularities enable researchers to adopt and analyze the multiple cognacy judgments for the three lexical forms. Table 7 illustrates the regularity comprising the V_1V_2 and V_2 types extracted from Table 5 that corresponds to the classification of *ciep*/(*cep*, *ceh*).

As shown in Table 7, the extracted regularities should improve or increase the information that was not captured by lexicostatistics. In the following section, we quantify the data of the 110 basic items used in Asai (1974) and the same data with the extracted regularities, that is, the former of which includes only Table 6, but the latter of which includes both Tables 6 and 7. Consequently, we demonstrate that the extracted and quantified regularities contribute to the clarification of Asai's (1974) original classification of the Ainu dialects numerically, compared to the classification results of the Ainu dialects obtained from the data without the regularities (i.e., 110 lexical items).

Table 7: The regularity (i.e., V_1V_2 and V_2 types) extracted from 19.fish.

	V_1V_2-type (*ciep*)	V_2-type (*cep*, *ceh*)
1_Yakumo	1	0
2_Oshamambe	1	0
3_Horobetsu	1	0
4_Biratori	0	1
5_Nukibetsu	0	1
6_Niikappu	0	1
7_Samani	0	1
8_Obihiro	0	1
9_Kushiro	0	1
10_Bihoro	0	1
11_Asahikawa	0	1
12_Nayoro	0	1
13_Soya	0	1
14_Ochiho	0	1
15_Tarantomari	0	1
16_Maoka	0	1
17_Shiraura	0	1
18_Raichishka	0	1
19_Nairo	0	1
20_North Kuril	0	1
21_Chitose	0	1

For example, Table 7 will be assigned to the different numbers in Table 6, and the assigned number in Table 7 will clearly change the disposition of certain dialects in the direction that Asai (1974) originally indicates from his lexicostatistical analysis, based only on the information of the 110 items.

4. Methods: How to Quantify "Regularity"

In the preceding sections, we have presented the methodology for extracting regularities and corresponding types from the basic items used in Asai (1974). The data with and

without these regularities and types are analyzed by homogeneity analysis that can quantify the two datasets, as introduced in the following subsections.

4.1 Why Homogeneity Analysis Should Be Applied in Quantifying Regularities

The main objective of this subsection is to explain why researchers should apply homogeneity analysis instead of correspondence analysis, which is often adopted in the realm of geolinguistics and dialectal segmentation theory, as discussed by Ono (2020b). Two linguistic points play a particularly significant role in quantifying the types in the extracted regularity, which can be achieved in homogeneity analysis but not in correspondence analysis: 1) the quantification of type should be reflected in the context of regularity; 2) the quantification of both type and regularity should be reflected in the configuration of languages and dialects.

4.2 The Quantification of Type Should be Reflected in the Context of Regularity

The information of type often depends on the context of regularity (e.g., geolinguistics). Table 8 shows a typical example of the tentative data of lexical items and forms, clearly illustrating the difference between correspondence and homogeneity analyses.

Table 8: The tentative data of lexical items and forms in Ono (2020b).

	Lexical item 1		Lexical item 2			Lexical item 3	
	Lexical form 1_1	Lexical form 1_2	Lexical form 2_1	Lexical form 2_2	Lexical form 2_3	Lexical form 3_1	Lexical form 3_2
1_Yakumo	1	0	1	0	0	1	0
2_Oshamambe	1	0	1	0	0	1	0
3_Horobetsu	1	0	1	0	0	1	0
4_Biratori	1	0	1	0	0	1	0
5_Nukibetsu	1	0	1	0	0	1	0
6_Niikappu	1	0	1	0	0	1	0
7_Samani	0	1	0	1	0	0	1
8_Obihiro	0	1	0	1	0	0	1
9_Kushiro	0	1	0	1	0	0	1
10_Bihoro	0	1	0	1	0	0	1
11_Asahikawa	1	0	1	0	0	1	0
12_Nayoro	1	0	1	0	0	1	0
13_Soya	0	1	0	1	0	0	1
14_Ochiho	1	0	0	0	1	1	0
15_Tarantomari	1	0	0	0	1	1	0
16_Maoka	1	0	0	0	1	1	0
17_Shiraura	1	0	0	0	1	1	0
18_Raichishka	1	0	0	0	1	1	0
19_Nairo	1	0	0	0	1	1	0
20_North Kuril	1	0	0	0	1	1	0
21_Chitose	1	0	0	0	1	1	0

In the case that it is the pattern of the lexical forms but not the pattern in the lexical items that researchers aim to focus on, the following three lexical forms (WF) of three lexical items (W) can be considered as identical: WF1_2 in W1, the WF2_2 in W2, and WF3_2 in W3[5] corresponding to gray cells in Table 8, respectively.

[5] Lexical item and lexical form correspond to "word" and "word form" in Asai (1974), respectively. The following part of this paper will abbreviate lexical item as W and lexical form as WF as consistent to terminologies in Asai (1974).

However, in another case where it is the pattern in the lexical items but not the pattern of the lexical forms that researchers aim to focus on, the three lexical forms can be considered as different. For example, the WF1_2 comprises the B-pattern in the ABA-pattern in W1 in geolinguistics, where the A-pattern is in the southwestern Hokkaido Ainu dialects (Nos. 1–6 in Figure 1) including the intermediate Asahikawa and Nayoro dialects (Nos. 11–12), and the North Kuril and Sakhalin Ainu dialects (Nos. 14–20), and the B-pattern is in the northeastern Hokkaido Ainu dialects (Nos. 7–10, 13). The same analysis applies to WF3_2 in W3; however, WF2_2 comprises a B-pattern (Nos. 7–10, 13) in the ABC-pattern of W2 in geolinguistics, where the A-pattern is in the southwestern Hokkaido Ainu dialects (Nos. 1–6), and the C-pattern is in the North Kuril and Sakhalin Ainu dialects (Nos. 14–20), with the exception of the Chitose dialect (No. 21).

In geolinguistics, the distinction between the ABA- and ABC-patterns is substantively significant; in the former, the A-pattern consisting of the periphery in the distribution is considered to reflect the older lexical form in the language or dialect. Thus, statistical methods that assign the same number to WF1_2, WF2_2, and WF3_2 in W1, W2, and W3, respectively, are inappropriate for the purpose in geolinguistics.

As the types (e.g., *iw*- and *uy*-types in the previous section) are extracted from several items (e.g., 179.pierce (stab): *ciw* (3, 4, 8, 9, 12–17, 19), *cuy* (10); 74.star: (*nociw* (1–9, 11, 12, 21), *noociw* (19), *nocuy* (9))) regarding regularity (e.g., the semivowel (glide) type), the type is necessarily considered in correspondence to another type in the regularity and is to be deprived of its substantive meaning without comparison to other types in the regularity. For example, the *uy*-type is invalid without correspondence to the *iw*-type, and these two types require the existence of the semivowel (glide) regularity.

Therefore, the analysis of Table 8 can also apply to the case of quantifying the types, replacing lexical forms as types and lexical items as regularities. As explained above, the quantification of the types corresponds to the case that it is the pattern in the regularities but not the pattern of the types that researchers aim to focus on. Thus, statistical methods quantifying the extracted types in our data are required for assigning different numbers to (WF1_2, WF3_2), and WF2_2 in (W1, W3), and W2, respectively, as shown in Table 9.

According to Ono (2020b), the left-hand side of Table 9 shows the assigned numbers obtained via correspondence analysis, while the right-hand side of Table 9 shows the assigned numbers obtained via homogeneity analysis. Since the assigned numbers differ between WF1_2 (or WF3_2) and WF2_2 in homogeneity analysis but are the same in WF1_2, WF2_2, and WF3_2 in correspondence analysis, homogeneity analysis is a more appropriate statistical method for quantifying the extracted types in the regularity.

Table 9: The results of quantification of the corresponding lexical forms using correspondence analysis and homogeneity analysis in Ono (2020b).

	Quantified scores			Quantified scores	
Correspondence analysis	1 dimension	2 dimension	Homogeneity analysis	1 dimension	2 dimension
Lexical form 1_1	-0.5976	0.0000	Lexical form 1_1	-0.0792	0.0000
Lexical form 1_2	1.6733	0.0000	Lexical form 1_2	0.2216	0.0000
Lexical form 2_1	-0.5976	1.7474	Lexical form 2_1	0.0000	-0.1336
Lexical form 2_2	1.6733	0.0000	Lexical form 2_2	0.0000	0.0000
Lexical form 2_3	-0.5976	-2.3299	Lexical form 2_3	0.0000	0.1782
Lexical form 3_1	-0.5976	0.0000	Lexical form 3_1	-0.0792	0.0000
Lexical form 3_2	1.6733	0.0000	Lexical form 3_2	0.2216	0.0000

4.3 The Quantification of Both Type and Regularity Should Be Reflected in the Configuration of Languages and Dialects

Another advantage of homogeneity analysis is that it allows researchers to quantify both type and regularity in such a way that the configuration of Ainu dialects can reflect the numbers assigned to lexical items and lexical forms, as shown in Table 10, where W1, W2, and W3 are the numbers assigned to each lexical item; WF1_1, WF1_2, WF2_1, WF2_2, WF2_3, WF3_1, and WF3_2 are the numbers assigned to each lexical form; and D1–21 are the numbers assigned to each dialect via homogeneity analysis, respectively. For example, WF1_1, WF1_2, WF2_1, WF2_2, WF2_3, WF3_1, and WF3_2 are calculated as -0.0792, 0.2216, 0.0000, 0.0000, 0.0000, -0.0792, and 0.2216, respectively, in the case of the first dimension in the right-hand side of Table 9.

Table 10: The quantification of the 21 Ainu dialects (i.e. D1–21) and the calculated configuration of each dialect based on the corresponding lexical items and forms using homogeneity analysis.

	Lexical item 1 (W1)		Lexical item 2 (W2)			Lexical item 3 (W3)		Calculated configuration of each dialect by quantified lexical forms and items	Configuration of each dialect quantified by homogeneity analysis
	Lexical form 1_1 (WF1_1)	Lexical form 1_2 (WF1_2)	Lexical form 2_1 (WF2_1)	Lexical form 2_2 (WF2_2)	Lexical form 2_3 (WF2_3)	Lexical form 3_1 (WF3_1)	Lexical form 3_2 (WF3_2)		
1_Yakumo	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D1
2_Oshamambe	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D2
3_Horobetsu	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D3
4_Biratori	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D4
5_Nukibetsu	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D5
6_Niikappu	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D6
7_Samani	0	WF1_2	0	WF2_2	0	0	WF3_2	W1*WF1_2+W2*WF2_2+W3*WF3_2	D7
8_Obihiro	0	WF1_2	0	WF2_2	0	0	WF3_2	W1*WF1_2+W2*WF2_2+W3*WF3_2	D8
9_Kushiro	0	WF1_2	0	WF2_2	0	0	WF3_2	W1*WF1_2+W2*WF2_2+W3*WF3_2	D9
10_Bihoro	0	WF1_2	0	WF2_2	0	0	WF3_2	W1*WF1_2+W2*WF2_2+W3*WF3_2	D10
11_Asahikawa	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D11
12_Nayoro	WF1_1	0	WF2_1	0	0	WF3_1	0	W1*WF1_1+W2*WF2_1+W3*WF3_1	D12
13_Soya	0	WF1_2	0	WF2_2	0	0	WF3_2	W1*WF1_2+W2*WF2_2+W3*WF3_2	D13
14_Ochiho	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D14
15_Tarantomari	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D15
16_Maoka	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D16
17_Shiraura	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D17
18_Raichishka	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D18
19_Nairo	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D19
20_North Kuril	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D20
21_Chitose	WF1_1	0	0	0	WF2_3	WF3_1	0	W1*WF1_1+W2*WF2_3+W3*WF3_1	D21

The estimated configuration of each dialect is calculated from W1, W2, W3, WF1_1, WF1_2, WF2_1, WF2_2, WF2_3, WF3_1, and WF3_2 as the sum of the product of the existing lexical form in each dialect, as presented in the penultimate column in Table 10, resulting in quantifying both type and regularity in such a way that the configuration of

the Ainu dialects can reflect the numbers assigned to the lexical items and forms. More precisely, homogeneity analysis numerically assigns W1, W2, W3, WF1_1, WF1_2, WF2_1, WF2_2, WF2_3, WF3_1, WF3_2, and D1–21 to minimize the discrepancy (i.e., sum of squared differences) between D1–21 and the sum of the product of the existing lexical items and forms in each dialect (e.g., to minimize the discrepancy between D1 and W1*WF1_1 + W2*WF2_1 + W3*WF3_1 = W1*-0.0792 + W2*0.0000 + W3*-0.0792 for the Yakumo dialect). Consequently, homogeneity analysis results in quantifying similar types within the context of the regularity to similar numbers, as well as a similar regularity to a similar number from the standpoint of mathematical and statistical views (Gifi 1990), achieving that the quantified configuration of the Ainu dialects is approximately reflected in sum of the product of the numbers assigned to lexical items and forms.

The numerical property of homogeneity analysis, whereby researchers can approximately visualize the relationship between the quantified types (and regularity) and the quantified configuration of each dialect, is also suitable for the purposes of this paper that aims to examine how the quantified types and regularity affect the classification of the Ainu dialects and which quantified types and regularity are relatively influential or not, as discussed in Section 5.

4.4 Homogeneity Analysis Revised for Quantifying Types and Regularities

The main objective of this subsection is to explain the motivation for revising the algorithm of homogeneity analysis (de Leeuw and Mair 2009) implemented in R language (R Core Team 2022) to appropriately quantify the data shown in Table 11.

In the tentative example on the left-hand side of Table 11, we assume that there are four types in regularity 1, and the Samani dialect is assumed to have all four types, for example. These data, which allow the dialect to contain multiple types in the regularity or multiple lexical forms in the item, are called multiple-answer data or multiple-choice-type data (e.g., Yamada and Nishisato 1993), while the data, which allow the dialect to contain a single type in the regularity or single lexical form in the item, are generally called item-category data.

Since the algorithm of homogeneity analysis (de Leeuw and Mair 2009) can only apply to item-category data and the advanced version of the algorithm (Mair and de Leeuw 2022) can apply to only limited case of multiple-answer data, we have revised the algorithm of homogeneity analysis for multiple answer data in this paper.

There was an issue regarding our implementation of the revised homogeneity analysis for the multiple-answer data, as shown in Table 11. On the right-hand side of Table 11, we transformed the tentative data of regularity 1 such that the sum of each row is equal to 1, consistent with the advanced version of the algorithm of homogeneity analysis (Mair and de Leeuw 2022). However, the right-hand side of Table 11 clearly shows a

contradiction, whereby the information regarding the presence of the type decreases to 0 as the number of types in the regularity increases in the dialect. For example, if the regularity 1 is assumed to increase one type and contain five types, and all types are observed in the Soya dialect, then the information of presence (1) is transformed as 1/5 = 0.2. Similarly, if the regularity 1 is assumed to increase 6 types and contain 10 types, and all types are observed in the Soya dialect, then the information of presence (1) is transformed as 1/10 = 0.1.

Table 11: The tentative data of regularity 1 and the transformed format of the data. Original data is assumed for our revised algorithm on homogeneity analysis (de Leeuw and Mair 2009).

Original data	Regularity 1 (tentative)				Transformed by homogeneity analysis	Regularity 1 (tentative)			
	type 1	type 2	type 3	type 4		type 1	type 2	type 3	type 4
1_Yakumo	1	0	0	1	1_Yakumo	0.5	0	0	0.5
2_Oshamambe	1	0	0	1	2_Oshamambe	0.5	0	0	0.5
3_Horobetsu	1	0	0	0	3_Horobetsu	1	0	0	0
4_Biratori	1	1	0	0	4_Biratori	0.5	0.5	0	0
5_Nukibetsu	1	1	0	0	5_Nukibetsu	0.5	0.5	0	0
6_Niikappu	1	1	0	0	6_Niikappu	0.5	0.5	0	0
7_Samani	1	1	1	1	7_Samani	0.25	0.25	0.25	0.25
8_Obihiro	1	0	1	0	8_Obihiro	0.5	0	0.5	0
9_Kushiro	1	0	1	0	9_Kushiro	0.5	0	0.5	0
10_Bihoro	1	0	1	0	10_Bihoro	0.5	0	0.5	0
11_Asahikawa	1	0	1	0	11_Asahikawa	0.5	0	0.5	0
12_Nayoro	1	0	1	0	12_Nayoro	0.5	0	0.5	0
13_Soya	1	1	1	1	13_Soya	0.25	0.25	0.25	0.25
14_Ochiho	0	1	1	0	14_Ochiho	0	0.5	0.5	0
15_Tarantomari	0	1	1	0	15_Tarantomari	0	0.5	0.5	0
16_Maoka	0	1	1	0	16_Maoka	0	0.5	0.5	0
17_Shiraura	0	1	1	0	17_Shiraura	0	0.5	0.5	0
18_Raichishka	0	1	1	0	18_Raichishka	0	0.5	0.5	0
19_Nairo	0	1	1	0	19_Nairo	0	0.5	0.5	0
20_North Kuril	0	1	1	0	20_North Kuril	0	0.5	0.5	0
21_Chitose	0	0	1	1	21_Chitose	0	0	0.5	0.5

Furthermore, as discussed in Section 3.3, the right-hand side of Table 11, transforming the original data as the sum of each row is equal to 1, can be regarded as more appropriate in situations where researchers possess some evidence from a theoretical perspective in linguistics that multiple types exist in the dialect, yet neither of these types can be observed in the linguistic materials in the dialect. For example, on the right-hand side of Table 11, it is more appropriate to consider the possession of evidence from a theoretical perspective in linguistics that only types 1 and 4 can exist in the Yakumo dialect, yet neither of the two types can be observed in the linguistic materials of the Yakumo dialect.

Thus, our revised algorithm does not transform our data, such that the sum of each row is equal to 1 in the dialect, but quantifies our original data, as shown on the left-hand side of Table 11. As explained in Fukazawa and Ono (2025), we extracted 13 regularities and 25 types from the 202 items in Asai (1974) and constructed Data 1, including only the 110 lexical items and the 398 lexical forms based on Asai's (1974) cognacy judgments; Data 2 are comprised of Data 1 and our extracted 13 regularities and 25 types.

5. Results

This section aims to clarify two points: first, how our extracted regularity affected the configuration of the Ainu dialects; second, which extracted regularity affected the configuration of the Ainu dialects from the results of our revised homogeneity analysis applied to Data 1 and 2.

5.1 How Our Extracted Type and Regularity Affected the Configuration of the Ainu Dialects

The main objective of this subsection is to explain the results of our statistical analyses that apply our revised homogeneity analysis to Data 1 and 2 in Section 4.

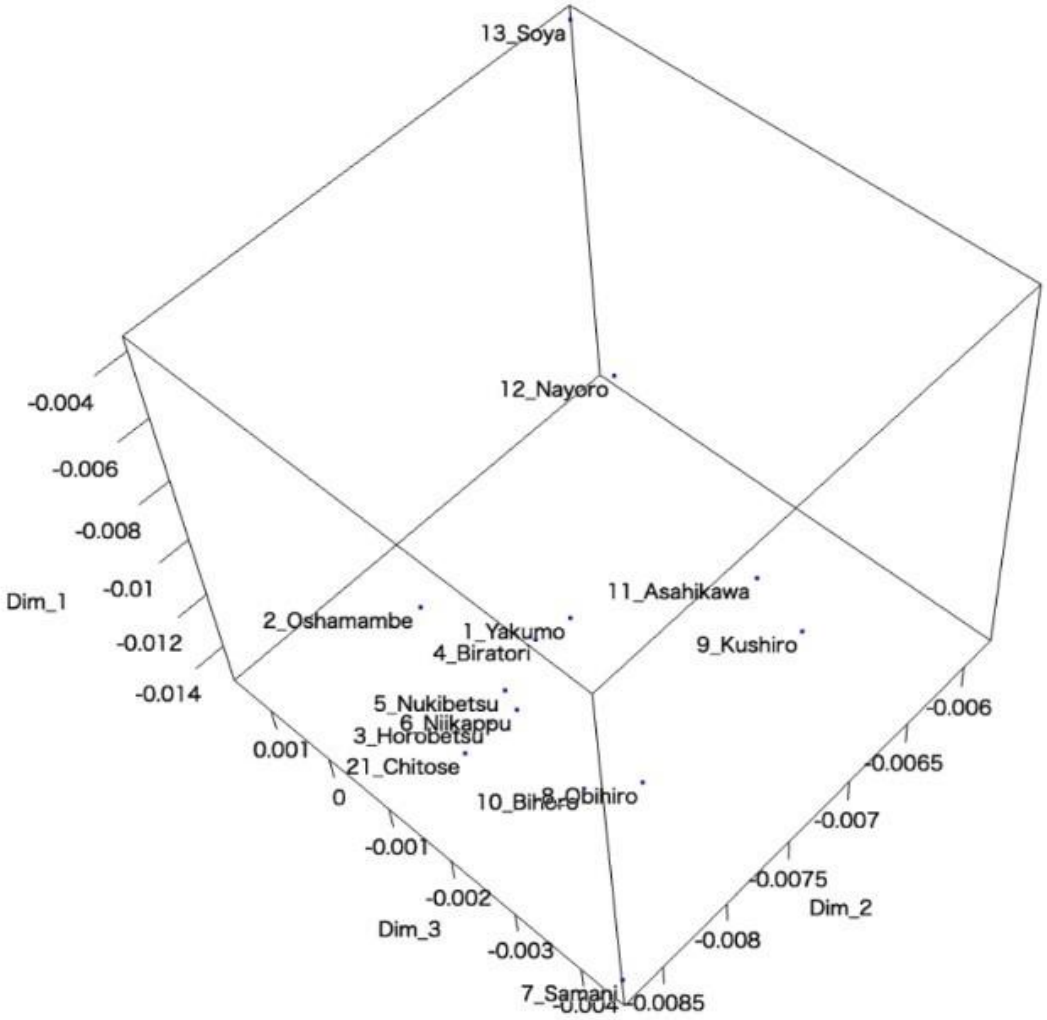


Figure 2: The configuration of the Ainu dialects obtained by homogeneity analysis applied to Data 1 (i.e., the 110 lexical items comprised of 398 lexical forms in Asai [1974]). Dim_1, Dim_2, and Dim_3 correspond to dimension 1, dimension 2, dimension 3, respectively.

We applied our revised homogeneity analysis to Data 1 and 2; quantified the lexical forms, items, types, regularities, and configurations of the 21 Ainu dialects in 3 dimensions; and visualized the quantification results in the 3-dimensional space.

Figures 2 and 3 show the configuration of the Ainu dialects obtained via the homogeneity analysis applied to Data 1 and 2, respectively, focusing on the Hokkaido Ainu dialect group in Asai (1974), that is, Nos. 1–13 and No. 21 in Figure 1.

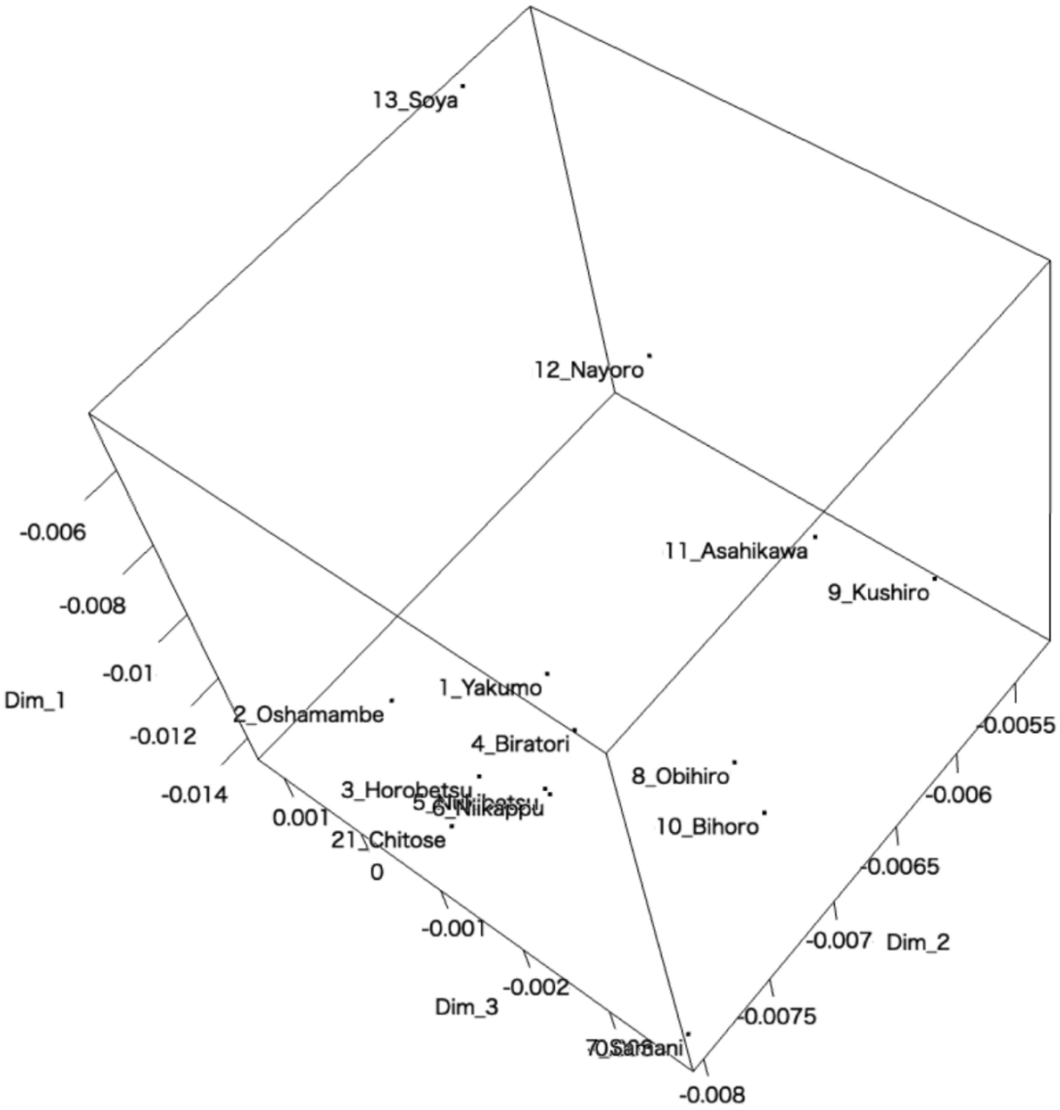
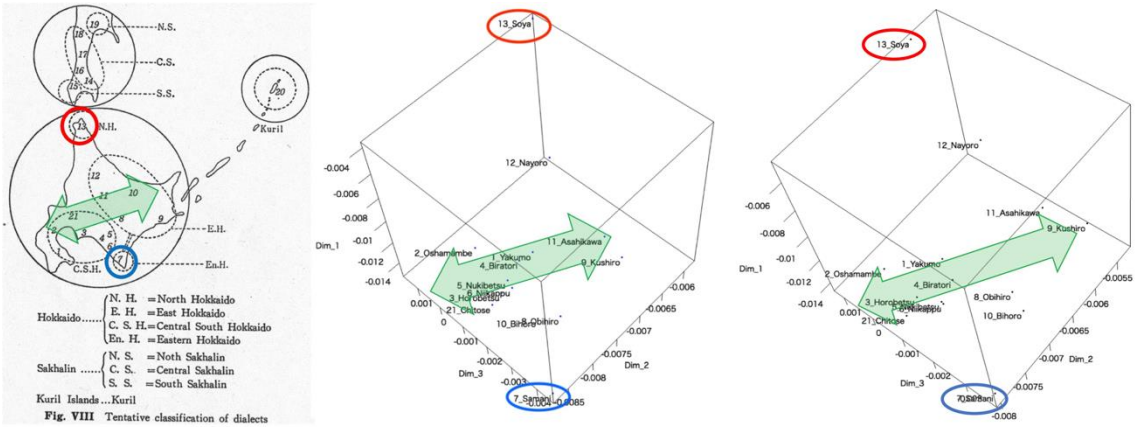


Figure 3: The configuration of the Ainu dialects obtained by homogeneity analysis applied to Data 2 (i.e., Data 1 and our extracted 13 regularities comprised of 25 types). Dim_1, Dim_2, and Dim 3 correspond to dimension 1, dimension 2, dimension 3, respectively.

Consequently, compared to Figure 1, Figure 2 shows more clearly the western and

eastern group axis comprising the Ainu dialects other than the Samani and Soya dialects of Hokkaido in terms of the northern and southern axis through the Samani and Soya dialects in Figure 4, the classification of the northeastern group in Figure 5, and the classification of the southwestern group in Figure 6.

Figure 4 shows the original classification of the 21 Ainu dialects in Asai (1974:100), with 2 distinctive points: 1) the northern and southern axis through the Samani and Soya dialects circled on the left-hand side of Figure 4; and 2) the western and eastern group axis comprising the Ainu dialects other than the Samani and Soya dialects of Hokkaido with the green bidirectional arrow on the left-hand side of Figure 4.



Asai (1974: 100) Data 1 (without regularity) Data 2 (with regularity)

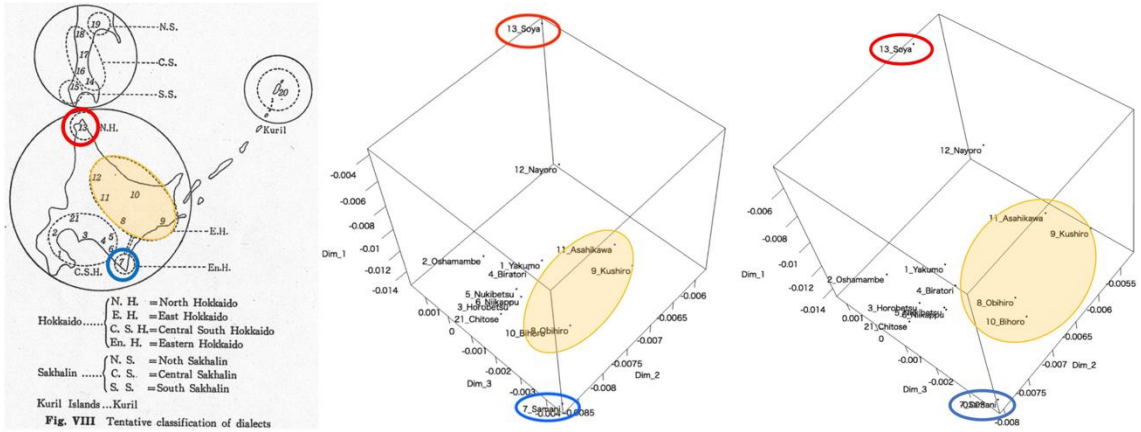
Figure 4: The northern and southern axis through the Samani and Soya dialects, the western and eastern group axis in Asai (1974), and our results obtained by homogeneity analysis. Left: Asai (1974: 100); Center: the configuration calculated from Data 1; Right: the configuration calculated from Data 2. The western and eastern group axis is illustrated with green bidirectional arrow.

Notably, the western and eastern group axis consisting of the Ainu dialects other than the Samani and Soya dialects of Hokkaido are more clearly elucidated in the configuration of the dialects using Data 2 than in that using Data 1. This clarified the characteristics of Asai's (1974) Ainu dialect classification using our extracted regularities in Data 2, as the result was not obtained in only original cognacy judgement of Data 1 (Ono and Fukazawa 2023, Fukazawa and Ono 2025).

Figure 5 and Figure 6 show the classification of the northeastern and southwestern Hokkaido Ainu dialect groups in Asai (1974), respectively, and our results obtained by homogeneity analysis. Moreover, the distinction among the northeastern and southeastern Hokkaido Ainu dialect groups is more clearly elucidated in the configuration of the dialects using Data 2 than in that using Data 1. This clarified the characteristics of Asai's (1974) Ainu dialect classification using our extracted regularities in Data 2, as the result was also not obtained in only original cognacy judgement of Data 1 (Ono and Fukazawa

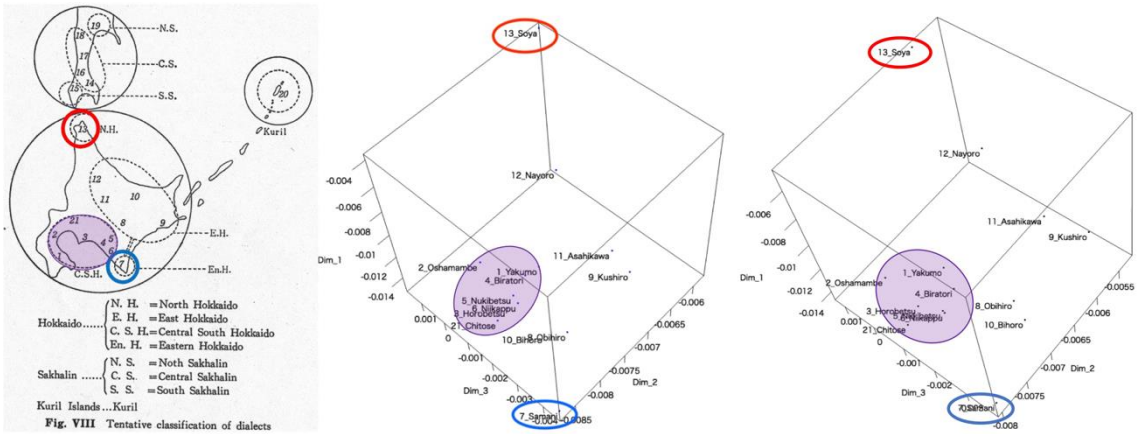
2023, Fukazawa and Ono 2025).

Thus, the results of the configuration using extracted regularities have led to clarifying the characteristics of Asai's (1974) Ainu dialect classification and suggested that the regularity information, analyzed from the viewpoints of Asai (1974), can be appropriately extracted and quantified by our proposed methods.



Asai (1974: 100) Data 1 (without regularity) Data 2 (with regularity)

Figure 5: The classification of the northeastern Hokkaido Ainu dialect group in Asai (1974) and our results obtained by our revised homogeneity analysis. Left: Asai's classification (Asai 1974: 100); Center: the configuration calculated from Data 1; Right: the configuration calculated from Data 2. The northeastern Hokkaido Ainu dialect group is surrounded by a circle.



Asai (1974: 100) Data 1 (without regularity) Data 2 (with regularity)

Figure 6: The classification of the southwestern group among the Ainu dialects in Asai (1974) and our results obtained by our revised homogeneity analysis. Left: Asai's classification (Asai 1974: 100); Center: the configuration calculated from Data 1; Right: the configuration calculated from Data 2. The southwestern Hokkaido Ainu dialect group is surrounded by a circle.

5.2 Which Extracted Type and Regularity Affected the Configuration of Ainu Dialects

In the previous subsection, we observed that our proposed methods appropriately extracted and quantified the regularity information in Asai (1974). The main objective of this subsection is to explain how the extracted types and regularities lead to the clarification of the characteristics of Asai's (1974) Ainu dialect classification and to show an example where the extracted regularity affects the configuration of certain Ainu dialects.

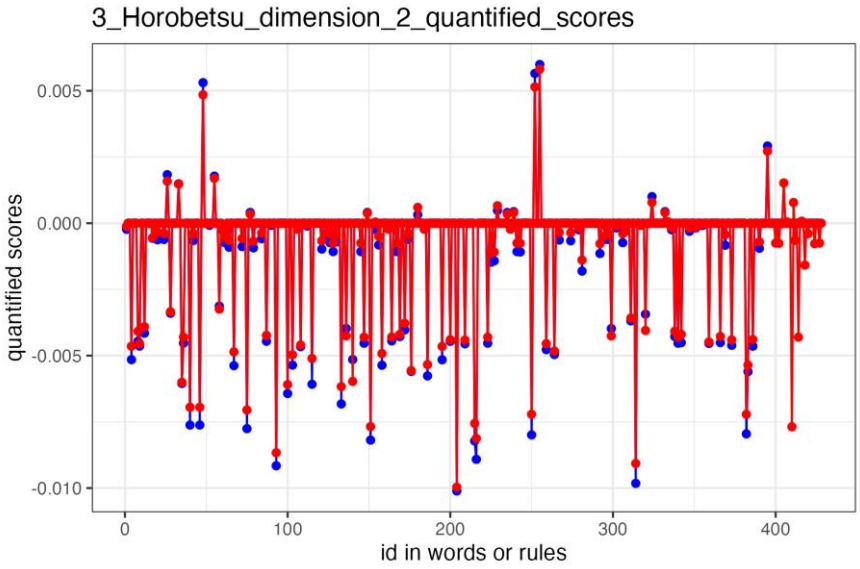


Figure 7: The assigned number to lexical forms and types of the second dimension in the Horobetsu dialect. Blue line: the quantified scores in Data 1. Red line: the quantified scores in Data 2. Note that the quantified score of 398 lexical forms in the 110 lexical items of Asai (1974) corresponds to the value of 1–398 IDs and the quantified score of 25 types in the 13 regularities corresponds to the value of 399–423 IDs in horizontal axis. IDs 410 and 411 correspond to V_1V_2-type and V_2 type from 19.fish, respectively.

Let us focus on the assigned numbers on the lexical forms in Data 1 and other assigned numbers on the lexical forms and types in Data 2, as shown in Figures A–L in the Supplementary Materials. As an example, Figure 7 shows the numbers assigned to lexical forms and types of the second dimension in the Horobetsu dialect, applying our revised homogeneity analysis to Data 1 and Data 2.

Figures A–C, D–F, and G–I show the assigned number to lexical forms and types of the first, second, and third dimensions in the Yakumo, Oshamambe, and Kushiro dialects, respectively. As the configuration of each dialect in Figures 2 and 3 is reflected in the sum of the quantified scores of lexical forms and types, as explained in Section 4, we can approximately consider from Figures A–I the extracted type that caused the clarification

of the characteristics of the classification in the Ainu dialects in Asai (1974), observing the numerical results that the pattern of lexical forms in blue line of Data 1 is approximately similar to the pattern of lexical forms and types in red line of Data 2 .

Furthermore, most of our extracted types showed negative values in Figures A–C, resulting in the changing of the Yakumo, Oshamambe, and Kushiro dialects in the negative direction in the first dimension from Figure 2 to Figure 3, while it did not clearly affect the change in the disposition of the Hokkaido Ainu dialects.

However, most of our extracted types showed negative values in the Yakumo and Oshamambe dialects, belonging to the southwestern Hokkaido Ainu dialect group in Asai (1974), as shown in Figures D and E. This resulted in changing the Yakumo and Oshamambe dialects in the negative direction in the second dimension from Figure 2 to Figure 3, clearly affecting the change in the disposition of the southwestern Hokkaido dialect group.

Conversely, most of our extracted types showed positive values in the Kushiro dialect, belonging to the northeastern Hokkaido Ainu dialect group in Asai (1974), as shown in Figure F. This resulted in a change in the Kushiro dialect into the positive direction in the second dimension from Figure 2 to Figure 3, and this clearly affected the change in the disposition of the northeastern Hokkaido dialect group. Moreover, some of our extracted types showed negative values, but others showed positive values in Figures G–I, resulting in the changing of the Yakumo, Oshamambe, and Kushiro dialects into both negative and positive directions in the third dimension, from Figure 2 to Figure 3. Therefore, further investigations are needed to determine which extracted type and regularity affect the configuration of Ainu dialects in the third dimension.

Thus, at least in the second dimension, we can identify that the quantified scores of our extracted types resulted in clarifying 1) the western and eastern group axis comprising the Ainu dialects other than the Samani and Soya dialects of Hokkaido in terms of the northern and southern axis through the Samani and Soya dialects in Figure 4, 2) the classifications of the northeastern group among the Hokkaido Ainu dialects in Figure 5, and 3) the classifications of the southwestern group among the Hokkaido Ainu dialects in Figure 6.

Finally, let us focus on Figures J–L, which correspond to the assigned number of lexical forms and types of the first, second, and third dimensions in the Horobetsu dialect. Figure K (i.e., Figure 7 in this paper) shows a greater negative value of -0.0077 in the V_1V_2 type extracted as *ciep*/(*cep*, *ceh*) from 19.fish, as shown in Table 7 in Section 3.4.

Figure K also reveals that the negative value of the V_1V_2 type clearly changes the disposition of the Horobetsu dialect in a negative direction, thereby evidently affecting the change in the Horobetsu dialect in the disposition of the southwestern Hokkaido dialect group. It is important to note that the lexical forms (*cep*, *ciep*) in Table 6 show a very small negative value of -0.0004, demonstrating that the extracted type in regularity,

not captured by lexicostatistics, clearly contributes to elucidating the characteristics of Asai's (1974) Ainu dialect classification with numerical evidence.

Therefore, our proposed methods enhanced the information in the basic vocabulary of the Ainu dialects in Asai (1974), which was not captured by lexicostatistics, and appropriately quantified types and regularities reflecting Asai's (1974) analysis of the Ainu dialects (cf., Fukazawa and Ono 2025), thus clarifying the characteristics of Asai's (1974) classification of the Ainu dialects.

6. Discussions and Conclusion

This final section discusses the significance of this paper from a methodological perspective. Our results suggest that our proposed methods can extract and quantify the information in linguistic materials that cannot be dealt with in the traditional framework of lexicostatistics. Since lexicostatistics does not allow for classifying lexical forms in multiple ways but only single way, our proposed method can extract the types in the regularity that are hidden by a single cognacy judgment, as explained in Section 3.

Thus, our proposed methods actually enhanced the information in the basic vocabulary of the Ainu dialects in Asai (1974), and the quantified types and regularities, which reflected Asai's analysis and cognacy judgement in the Ainu dialects, resulted in clarifying the characteristics in the classification of the Ainu dialects by Asai.

Furthermore, the scope of our proposed methods can transcend lexicostatistics. As reported by Swadesh (1955), lexicostatistics has fundamental issues that entail what items should be included in and excluded from the basic vocabulary; the demarcation problems regarding the basic vocabulary have restricted the scope of lexicostatistics to lexical forms of the basic vocabulary.

However, our proposed methods are not confined to the lexical forms in the basic vocabulary; they are also applicable to other vocabulary items that are not included in the basic vocabulary but that can be regarded as rich resources in language and dialect. Consequently, the proposed methods will enable researchers to facilitate the exploration of these resources for linguistic research.

Note that the proposed methods can be applied to any type of information, including the types and regularities thereof. In other words, these methods have the potential to enable researchers to extract and quantify the types and regularities across the present subdivisions of linguistics, phonology, morphology, and syntax, among others. For example, the Saru and Chitose dialects of Ainu are different from other Hokkaido dialects in a few ways. They have different person marking system, different morphemes between interrogative and infinitive forms and the existence of a singular-plural distinction in the certain person (cf., Nakagawa and Fukazawa 2022).

Remarkably, the extracted and quantified information will elucidate hidden structures in multidimensional visualization, whereby the combination of the calculated dimensions

can reflect the divergence and geographical information of the language and dialect, as shown in our analyses of the Ainu dialects.

Moreover, while quantification theory in statistics has mainly focused on presence (1) or absence (0) in the data, the types in regularity contain more complex structures that cannot be dealt with in the present quantification theory. For example, there are sometimes historical directions between types in regularity, as stated by Fukazawa and Ono (2025). The types and regularities were extracted as hV_1 > $'V_1$ from 123.wash: (*huraye* (1–7, 13–19, 21), *uraye* (8–12)), corresponding to *huraye* > *uraye*, from the historical perspective of the Ainu language. Conversely, other types and regularity were extracted as $'V_1$ > h-+V_1 from 78.sand: (*ota* (1–4, 6, 8–21), *hota* (7)), corresponding to *ota* > *hota* from the historical perspective of the Ainu language.

However, our proposed methods did not consider directional information between types in this study. Therefore, further statistical investigations can contribute to not only the development of linguistics research but also quantification theory in statistics.

In conclusion, the development of methodology is not necessarily confined to a meticulous analysis of the subject matter. As demonstrated in this study, it is possible to expand the range of subjects and information that can be addressed in the realm by pursuing this methodology. In the future perspectives, our proposed methods also have the potential to contribute to the quantification of vocabulary items and the classification of the other languages (e.g., Native North Americas), which should have more complex phonological and morphological characteristics than the Ainu language. We end this paper with the hope that interdisciplinary research will prove to be invaluable in our direction in the future.

Acknowledgments

Parts of this paper were presented at the 7th Annual Conference of The Japan Association of Northern Language Studies, and we are verry grateful to the conference organizers and participants. Also, we would like to express our deepest gratitude to editor and two highly conscientious reviewers for their constructive and invaluable comments; all errors are of course our own.

References

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

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

de Leeuw, Jan and Patrick Mair (2009) Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, 31, 1–21.

Fukazawa, Mika and Yohei Ono (2024) Hogen kyokai saiko: ainugo no hogen kyokai wo

rei toshite [Revisiting the concept of dialectal boundary: an exercise on the Ainu dialects]. *Northern Language Studies*, 14: 155–176.

Fukazawa, Mika and Yohei Ono (2025) Kisogoi niyoru hogen bunrui no shomondai: Asai (1974) no ainugo hogen deta wo saidaigen katsuyo surutame ni [Challenges of dialect classification in extracting maximum information: from the basic vocabulary of Ainu in Asai (1974)]. *Northern Language Studies*, 15.

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

Gifi, Albert (1990) *Nonlinear multivariate analysis*. New York: John Wiley and Sons.

Mair, Patrick and Jan de Leeuw (2022) Gifi: Multivariate analysis with optimal scaling. R-package version 0.4-0. https://cran.r-project.org/package=Gifi

Hattori, Shiro and Mashiho Chiri (1960) Ainugo shohogen no kisogoi tokeigakuteki kenkyū [A lexicostatistic study on Ainu dialects]. *Kikan Minzokugaku Kenkyū [The Japanese Journal of Ethnology]*, 24(4), 307–342.

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

Murayama, Shichiro (1971) *Kita Chishima Ainu-go [Ainu language of Northern Kuril Islands].* Tokyo: Yoshikawa-Kobun-Kan.

Nakagawa, Hiroshi and Mika Fukazawa (2022) Hokkaido Ainu dialects: towards a classification of Ainu dialects. In: Bugaeva, Anna (ed.) *Handbook of the Ainu language*, 253–328. Berlin/Boston: Mouton De Gruyter.

Ono, Yohei (2020a) Some remarks on cognacy judgments of Ainu dialects: on Asai (1974). *Journal of the Center for Northern Humanities*, 13: 37–57.

Ono, Yohei (2020b) Gengogaku niokeru “suryoka” wo saiko suru: hoppo shogengo no gengochirigaku to gengoruikeiron no taihiwo tsujite [Reconsideration of “quantification” in linguistics: through linguistic geography and language typology in northern language studies]. Presentation at the 3rd Annual Conference of The Japan Association of Northern Language Studies. Abashiri (Online).

Ono, Yohei and Mika Fukazawa (2023) Ainugo shohogen no gokei no ruiji nikansuru kiso deta no fukugen: ronbun ni kaki kirenakatta kenkyusha no handan shikoni semaru [Reconstruction of original data on similarity between lexical forms in Ainu dialects: approaching the researcher’s unwritten judgment and thoughts]. *Northern Language Studies*, 13: 213–245.

Ono, Yohei and Mika Fukazawa (2024) Hikakufukano datta ainugo hogenbunrui: tokeiteki hogenbunrui wo ruijido no tenkara saiko suru [(Un)comparable classifications of Ainu dialects: Reconsidering statistical dialect classification from the similarity judgments]. *Journal of Ainu and Indigenous Studies*, 4: 93–126.

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Swadesh, Morris (1955) Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 24, 121–137.
Torii, Ryūzo (1903) *Chishima Ainu [Kuril Ainu]*. Tokyo: Yoshikawa-Kobun-Kan.
Yamada, Fumiyasu and Shizuhiko Nishisato (1993) Sotsui shakudoho nikansuru ikutsuka no tokusei [Several mathematical properties of dual scaling as applied to dichotomous item-category data]. *Kodo Keiryogaku [The Japanese Journal of Behaviormetrics]*, 20(1): 56–63.

Summary

The information of "regularity," so-called phonological correspondences, morphological and grammatical rules, is always present in the data of languages and dialects, but it cannot necessarily be analyzed in a numerical way that is applied to lexical items and forms in previous linguistics research (e.g., lexicostatistics). Therefore, the shortcoming of the methodologies can potentially prevent previous studies from approaching linguistic issues due to a lack of appropriately quantified information of "regularity." This paper attempts to provide some possibility regarding how researchers can extract and quantify the regularities in languages and dialects, using the data of Ainu dialects in Asai (1974).

We propose principles for extracting the type and regularity from linguistic materials and statistical quantification methods that quantify the type and regularity more appropriately from the viewpoints of both linguistics and statistics.

Consequently, the classification result obtained by incorporating the quantified types and regularities further clearly shows the western and eastern group axis comprising the Ainu dialects other than the Samani and Soya dialects of Hokkaido in terms of the northern and southern axis through the Samani and Soya dialects, the classification among the northeastern Hokkaido Ainu dialect group, and the classification among the southwestern Hokkaido Ainu dialect group.

Our results have clarified the characteristics of Asai's (1974) Ainu dialect classification and suggested that regularity information, excluded from Asai's (1974) classification as regularity while reflecting his viewpoints, can be appropriately extracted and quantified by our proposed methods.

Notably, our proposed methods can be applied to any type of information, including the types and regularities thereof. In other words, these methods have the potential to enable researchers to extract and quantify the types and regularities across the present subdivisions of linguistics, phonology, morphology, and syntax, among others. The application of these methods can also facilitate the quantification of vocabulary items and the classification of the other languages (e.g., Native North Americas), which should have more complex phonological and morphological characteristics than the Ainu language.

（mathematical.humanities@hotmail.com / mk.fukazawa@hotmail.co.jp）