Northern Language Studies 16: 121-150 (Japan Association of NoLS, 2026)

From Classification to History:

Examining New Lexicostatistical Methodologies

in Using the Data in Ainu Dialects

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

Keywords: Ainu language, 3D visualization, data preparation, historical divergence, language contact

1. Introduction

Lexical-item data in languages or dialects should contain not only information about lexical forms but also linguistic (phonological or morphological) regularities. Previous studies (Fukazawa and Ono 2025, Ono and Fukazawa 2025) reported that adding information about regularities to Asai's (1974) lexical-item data, which have been excluded from Asai's (1974) statistical analysis (Ono and Fukazawa 2023, 2024), enabled more precise visualization of Asai's (1974) classifications of Ainu dialects by maximizing the information captured from lexical-item data.

Although the original scope of lexicostatistics encompassed the historical investigation of languages or dialects (Swadesh 1955), methodological limitations—specifically data-preparation methods described in this study—have restricted its practical scope to mere "classification."

Consequently, whether the information can be extracted from Asai's documents in Ainu dialectology to visualize historical insights remains unclear. In other words, the fundamental question is whether lexicostatistics can shift its focus from classification back to history.

Thus, this study primarily aimed to propose a new data-preparation method in lexicostatistics and to verify its validity, specifically by exploring whether our proposed method can potentially reveal historical divergence or language contact in Ainu dialects—aspects that may have remained obscured due to methodological limitations in current preparation methods.

The remainder of this study is organized as follows: Section 2 provides our new lexicostatistical methodologies, taking Asai's (1974) lexicostatistical documents as an example. The first part focuses on data-preparation methods consisting of multiple cognacy judgements and two datasets—the lexical dataset and a dataset combining lexical

and regularity information—and the second part focuses on a quantification method that assigns numerical values to the symbols in the two datasets.

Section 3 presents three-dimensional (3D) visualizations of Ainu dialects derived from the application of a quantification method to two datasets. Section 4 depicts the methodological significance of our proposed method from both linguistic and statistical perspectives.

2. Methods

This study develops new lexicostatistical methodologies, using the lexicostatistical documents in Asai (1974) as a case study.

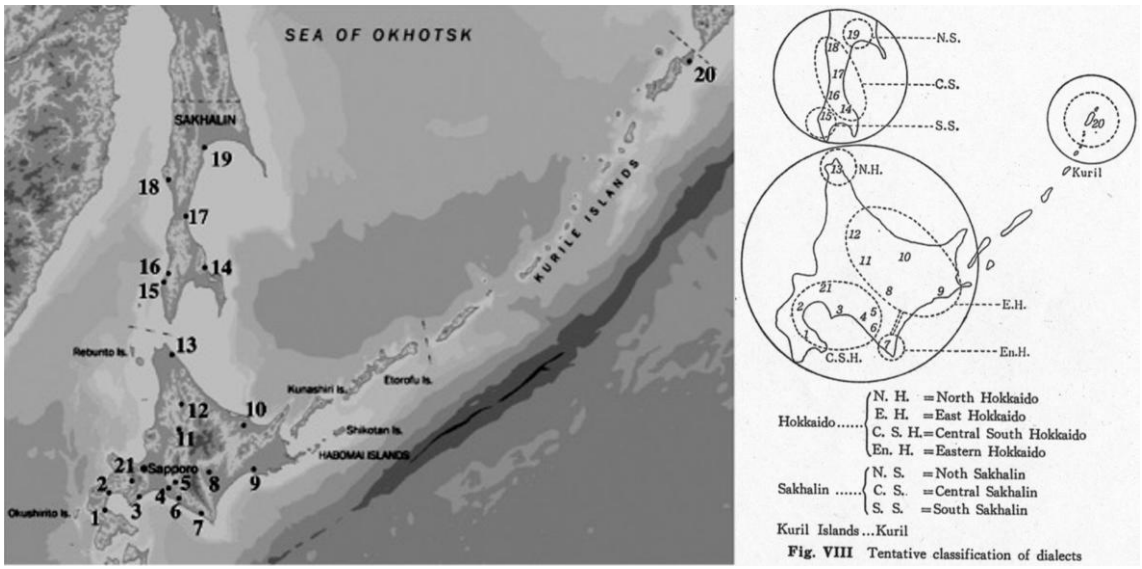

Figure 1. Map of Ainu dialects (Geospatial Information Authority of Japan, 2019).[1] Left: Map of Ainu dialects (Geospatial Information Authority of Japan, 2019). Note: 1, Yakumo; 2, Oshamambe; 3, Horobetsu; 4, Biratori; 5, Nukibetsu; 6, Niikappu; 7, Samani; 8, Obihiro; 9, Kushiro; 10, Bihoro; 11, Asahikawa; 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 Ainu dialects (Asai 1974: 100).

Two lexicostatistical studies have classified Ainu dialects that significantly influence the current Ainu linguistics, as discussed by Ono (2020: 37). The first, compiled by Hattori and Chiri (1960), covers 19 dialects (Nos. 1–19 in Figure 1) using a 200-word list adapted from Swadesh (1955) (cf. Ono 2020, Ono and Fukazawa 2023, Ono and Fukazawa 2024, 2025). Hattori and Chiri (1960) investigated linguistic documents from that time, as many dialects were on the verge of disappearing. The second study, Asai

[1] The area where the Ainu language was spoken is located in the southern part of Sakhalin Island, and the area is conventionally called "Sakhalin" in Hattori and Chiri (1960) and Asai (1974).

(1974), serves as an expansion and revision of the first. Asai (1974) not only corrected the data based on Hattori (1964) but also updated the lexical documents for the Obihiro, Kushiro, and Asahikawa dialects (Nos. 8, 9, and 11 in Figure 1) based on his own fieldwork. Moreover, he expanded the scope of the survey by adding the Chitose dialect (No. 21 in Figure 1) obtained from informants and the North Kuril dialect (No. 20 in Figure 1) derived from historical documents by Torii (1903), Murayama (1971), and Pinart's vocabulary manuscripts (Asai 1974: Appendix). The history of lexicostatistical studies of Ainu dialects is explained in detail in Section 2 on the supplementary materials.

2.1 A General Framework of Lexicostatistical Methodologies

Figure 2 shows the flowchart of the lexicostatistical methodologies. As a first step, lexicostatistical documents are transformed into extracted data through data preparation. The lexicostatistical methodologies consist of three components: data preparation, quantification method, and visualization. This paper discusses previous problems and our proposed solutions for each component.

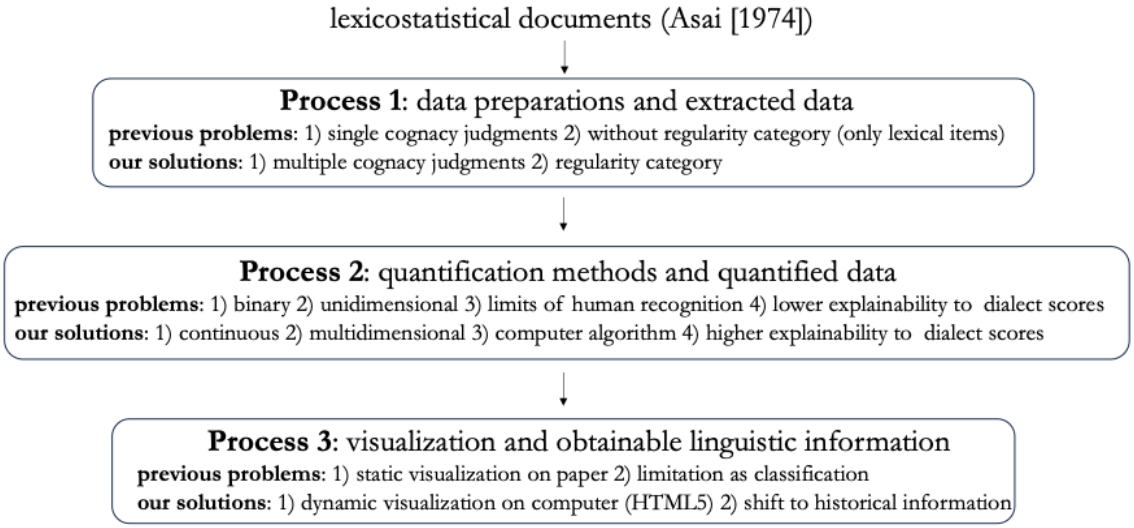

Figure 2. Flowchart of lexicostatistical methodologies.

Table 1 provides a typical example of Asai's (1974) lexicostatistical documents on Ainu dialects. For instance, the informant for the Yakumo dialect (No. 1)[2] provided the lexical form *nen* for the meaning 'who' (Item 6 of the 200 basic vocabulary list), while the informant for the Kushiro dialect (No. 9) provided both *ruyanpe* and *ruwanpe* for the meaning 'rain (the rain)'.

Several lexical forms in Table 1 can be considered to share the same origin from the

[2] In following explanations, the number with No. corresponds to the number of places in Figure 1.

perspective of Ainu linguistics: *hunna* and *hunat* for ‘who’; *ruyanpe* and *ruwanpe* for ‘rain (the rain)’; *apto*, *ahto*, and *atto* for ‘rain (the rain)’; *nekona* and *nekon* for ‘how’; *mak*, *manak*, and *makanak* for ‘how’; *hekaci* and *hekattar* for ‘child’; *ekaci* and *ekattar* for ‘child’; and *rarak*, *raarak*, *raarah*, and *taarak* for ‘smooth’.

Table 1. Example of lexicostatistical documents in Asai (1974).

	6. who	76. rain (the rain)	104. how	151. child	167. smooth
1_Yakumo	nen	weni	nekona	hekaci	rarak
2_Oshamambe	nen	weni	nekona	hekattar	rarak
3_Horobetsu	nen	apto	nekona	hekattar	rarak
4_Biratori	hunna	apto	mak	hekattar	rarak
5_Nukibetsu	hunna	apto	mak	hekattar	rarak
6_Niikappu	hunna	apto	mak	hekattar	rarak
7_Samani	nen	ruyanpe	nekon	hekattar	rarak
8_Obihiro	nen	ruyanpe	nekon	ekaci	testek
9_Kushiro	nen	ruyanpe, ruwanpe	nekona, nekon	ekaci	testek
10_Bihoro	nen	ruwanpe	nekona, nekon	ekaci	testek
11_Asahikawa	nen	ruyanpe	nekon	hekattar, ekaci, ekattar	rarak
12_Nayoro	nen	ruyanpe	nekon	ekattar	testek
13_Soya	nen	ruyanpe	nekon	hekaci	rarak
14_Ochiho	naata	ahto	temana	poo	raarah
15_Tarantomari	naata	atto	temana	hekaci	raarak
16_Maoka	naata	ahto	temana	hekaci	raarah
17_Shiraura	naata	ahto	temana	hekaci	raarah
18_Raichishka	naata	ahto	temana	hekaci	raarah
19_Nairo	naata	atto	temana	hekaci	taarak
20_North Kuril	hunat	sirun	uiman	poo	rarak
21_Chitose	hunna	apto	manak, makanak	hekaci	rarak

Table 2. Example of extracted data obtained through data preparation applied to Table 1. Uppercase symbols correspond to the summarization of lexical forms.

	6. who	76. rain (the rain)	104. how	151. child	167. smooth
1_Yakumo	nen	weni	NEKONA	HEKACI	RARAK
2_Oshamambe	nen	weni	NEKONA	HEKACI	RARAK
3_Horobetsu	nen	APTO	NEKONA	HEKACI	RARAK
4_Biratori	HUNNA	APTO	MAK	HEKACI	RARAK
5_Nukibetsu	HUNNA	APTO	MAK	HEKACI	RARAK
6_Niikappu	HUNNA	APTO	MAK	HEKACI	RARAK
7_Samani	nen	RUYANPE	NEKONA	HEKACI	RARAK
8_Obihiro	nen	RUYANPE	NEKONA	EKACI	testek
9_Kushiro	nen	RUYANPE	NEKONA	EKACI	testek
10_Bihoro	nen	RUYANPE	NEKONA	EKACI	testek
11_Asahikawa	nen	RUYANPE	NEKONA	HEKACI, EKACI	RARAK
12_Nayoro	nen	RUYANPE	NEKONA	EKACI	testek
13_Soya	nen	RUYANPE	NEKONA	HEKACI	RARAK
14_Ochiho	naata	APTO	temana	poo	RARAK
15_Tarantomari	naata	APTO	temana	HEKACI	RARAK
16_Maoka	naata	APTO	temana	HEKACI	RARAK
17_Shiraura	naata	APTO	temana	HEKACI	RARAK
18_Raichishka	naata	APTO	temana	HEKACI	RARAK
19_Nairo	naata	APTO	temana	HEKACI	RARAK
20_North Kuril	HUNNA	sirun	uiman	poo	RARAK
21_Chitose	HUNNA	APTO	MAK	HEKACI	RARAK

Thus, Table 1 can be transformed into Table 2, which is referred to in this paper as "extracted data," by applying prior cognacy judgments as illustrative examples. For instance, HUNNA represents the summarized lexical form encompassing *hunna* and *hunat*. The procedure shown as Process 1 in Figure 2, in which lexicostatistical documents are transformed into extracted data, is termed "data preparation" in this paper.

Table 3. Example of quantified data obtained through a quantification method applied to Table 1. As an example, the presence of lexical forms or their summarized lexical forms is coded as 1, and their absence as 0.

	6. who			76. rain (the rain)				104. how				151. child			167. smooth	
	nen	HUNNA	naata	weni	APTO	RUYANPE	sirun	NEKONA	MAK	temana	uiman	HEKACI	EKACI	poo	RARAK	testek
1_Yakumo	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
2_Oshamambe	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
3_Horobetsu	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0
4_Biratori	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0
5_Nukibetsu	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0
6_Niikappu	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0
7_Samani	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0
8_Obihiro	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1
9_Kushiro	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1
10_Bihoro	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1
11_Asahikawa	1	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0
12_Nayoro	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1
13_Soya	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0
14_Ochiho	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	0
15_Tarantomari	0	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0
16_Maoka	0	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0
17_Shiraura	0	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0
18_Raichishka	0	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0
19_Nairo	0	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0
20_North Kuril	0	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0
21_Chitose	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0

Table 4. Example of visualization applied to Table 3. As an example, sorting operations were applied to Table 3 according to similar patterns in the distribution of lexical forms across neighboring columns.

	HUNNA	MAK	APTO	weni	nen	NEKONA	HEKACI	RARAK	RUYANPE	EKACI	testek	naata	temana	poo	sirun	uiman
1_Yakumo	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0
2_Oshamambe	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0
3_Horobetsu	0	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0
4_Biratori	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0
5_Nukibetsu	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0
6_Niikappu	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0
7_Samani	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0
8_Obihiro	0	0	0	0	1	1	0	0	1	1	1	0	0	0	0	0
9_Kushiro	0	0	0	0	1	1	0	0	1	1	1	0	0	0	0	0
10_Bihoro	0	0	0	0	1	1	0	0	1	1	1	0	0	0	0	0
11_Asahikawa	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0
12_Nayoro	0	0	0	0	1	1	0	0	1	1	1	0	0	0	0	0
13_Soya	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0
14_Ochiho	0	0	1	0	0	0	0	1	0	0	0	1	1	1	0	0
15_Tarantomari	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0
16_Maoka	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0
17_Shiraura	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0
18_Raichishka	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0
19_Nairo	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0
20_North Kuril	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1
21_Chitose	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0

As a next step, to extract linguistic information from Table 2 (e.g., through counting), it is necessary to assign numerical values to each lexical form in the dialects. Table 3 provides a simple example in which a value of 1 is assigned to the presence of a lexical

form or summarized lexical form, and a value of 0 is assigned to its absence. Table 3 is referred to as "quantified data," and the procedure shown as Process 2 in Figure 2, in which extracted data are transformed into quantified data, is termed "quantification method" in this paper.

Because Table 3 consists solely of numerical values, it remains difficult for researchers to directly recognize the linguistic information it contains. Accordingly, Table 3 must be further transformed to make such information visually accessible. Table 4 presents a simple example in which the columns of Table 3 are sorted according to similarities in the distribution patterns of lexical forms across neighboring dialects. These patterns are referred to as "obtainable linguistic information," and the procedure shown as Process 3 in Figure 2, in which quantified data are transformed into obtainable linguistic information, is termed "visualization" in this paper.

However, each of these three procedures—data preparation, quantification method, and visualization—contains inherent methodological problems, as discussed in the following subsection.

2.2 Data Preparations and Extracted Data

This subsection primarily explains our proposed data preparation, taking another example of lexical forms corresponding to '19. fish', '76. rain' (the rain), '148. ice' and '167. smooth' on Table 5 as an example.

2.2.1 Methodological Problems in Data Preparations

Two methodological problems were reported in previous studies: the lack of multiple cognacy judgments and the lack of a "regularity" category. First, let us focus on the absence of multiple cognacy judgments in previous data-preparation methods.

Table 5. Another example of lexical forms in Asai (1974).

	19. fish	76. rain (the rain)	148. ice	167. smooth
1_Yakumo	ciep	weni	konru	rarak
2_Oshamambe	ciep	weni	konru	rarak
3_Horobetsu	cep, ciep	apto	konru	rarak
4_Biratori	cep	apto	konru	rarak
5_Nukibetsu	cep	apto	konru	rarak
6_Niikappu	cep	apto	konru	rarak
7_Samani	cep	ruyanpe	konru	rarak
8_Obihiro	cep	ruyanpe	konru	testek
9_Kushiro	cep	ruyanpe, ruwanpe	konru	testek
10_Bihoro	cep	ruwanpe	konru	testek
11_Asahikawa	cep	ruyanpe	konru	rarak
12_Nayoro	cep	ruyanpe	konru, rup	testek
13_Soya	cep	ruyanpe	rup	rarak
14_Ochiho	ceh	ahto	ruh	raarah
15_Tarantomari	cep	atto	rup	raarak
16_Maoka	ceh	ahto	ruh	raarah
17_Shiraura	ceh	ahto	ruh	raarah
18_Raichishka	ceh	ahto	ruh	raarah
19_Nairo	cep	atto	tup	taarak
20_North Kuril	cep	sirun	konru	rarak
21_Chitose	cep, ciep	apto	konru	rarak

2.2.1.1 Absence of Multiple Cognacy Judgments

Until recent studies (Fukazawa and Ono 2025; Ono and Fukazawa 2025), data-preparation methods in lexicostatistics did not accommodate multiple cognacy judgments within a single lexical item. This limitation restricted the extracted information to a data format permitting only a single

cognacy judgment per lexical item.

As shown in Table 5, in Asai (1974), the lexical item ‘fish’ consists of three lexical forms: *cep* (3–13, 15, 19–21), *ciep* (1–3), and *ceh* (14, 16–18). Note that the numbers in parentheses following each lexical form correspond to the place number in Figure 1. In contrast, current data-preparation methods require the logical selection of one of the following five cognacy judgments: *cep*/*ciep*/*ceh*, *cep*/(*ciep*, *ceh*), *ciep*/(*cep*, *ceh*), *ceh*/(*cep*, *ciep*), or (*cep*, *ciep*, *ceh*).

Table 6. Example of the resulting quantified data obtained from the extracted data of Table 5.

	19. fish			76. rain (the rain)						148. ice				167. smooth				
	ciep	cep	ceh	weni	apto	RUYANPE	ahto	atto	sirun	konru	rup	ruh	tup	rarak	raarak	raarah	taarak	testek
1_Yakumo	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
2_Oshamambe	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
3_Horobetsu	1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
4_Biratori	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
5_Nukibetsu	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
6_Niikappu	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
7_Samani	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
8_Obihiro	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
9_Kushiro	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
10_Bihoro	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
11_Asahikawa	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
12_Nayoro	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1
13_Soya	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0
14_Ochiho	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
15_Tarantomari	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
16_Maoka	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
17_Shiraura	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
18_Raichishka	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
19_Nairo	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0
20_North Kuril	0	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0
21_Chitose	1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0

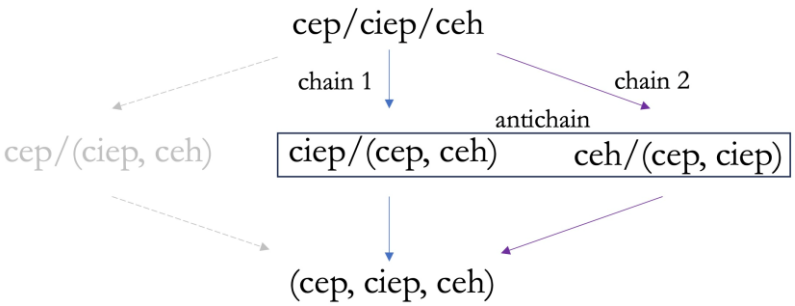

Figure 3. Chains and antichain in the lattice relationship among cognacy judgments in ‘fish.’ The directional arrow corresponds to the inclusion relationship of cognacy judgment shown as < in the main body of this paper. The left chain is highlighted in gray because the cognacy judgment *cep*/(*ciep, ceh*) is not admissible from the phonological viewpoint of Ainu language.

For example, Table 6 presents quantified data obtained from the extracted data, in which *cep* and *ciep* are judged to be noncognate, as are *cep* and *ceh*.

Mathematically, these relationships constitute a lattice structure (Grätzer 1998), as

illustrated in Figure 3. The theoretical point is the presence of two inclusion relationships regarding cognacy judgments (i.e., chains in lattice theory):

cep/*ciep*/*ceh* < *ciep*/(*cep*, *ceh*) < (*cep*, *ciep*, *ceh*) (chain 1)

cep/*ciep*/*ceh* < *ceh*/(*cep*, *ciep*) < (*cep*, *ciep*, *ceh*) (chain 2)

where the symbol < denotes that the left cognacy judgment (e.g., *cep*/*ciep*/*ceh)* is finer than the cognacy judgment on the right side. Furthermore, no inclusion relationship holds between *ciep/*(*cep*, *ceh*) and *ceh*/(*cep*, *ciep*), that is, neither *ciep*/(*cep*, *ceh*) < *ceh*/(*cep*, *ciep*) nor *ceh*/(*cep*, *ciep*) < *ciep*/(*cep*, *ceh*) is true. Thus, these two classifications constitute an antichain in the lattice structure of 'fish.'

The presence of these two antichain elements (i.e., *ciep*/(*cep*, *ceh*) and *ceh*/(*cep*, *ciep*)) indicates the need to allow multiple cognacy judgments for a single lexical item to maximize the extracted information. Consequently, previous lexicostatistical studies were restricted to determining which inclusion relationship (chain) to follow and at which level to fix their cognacy judgments, mainly because they adopted a single judgment for each lexical item.

For example, under the constraint of a single cognacy judgment, it is necessary to choose between two alternatives. One option is to include only chain 1 and select a single cognacy judgment within that chain (e.g., *cep/ciep/ceh*, *ciep/*(*cep, ceh*), or (*cep, ciep, ceh*)), in which case the information represented by *ceh/*(*cep, ciep*) in chain 2 cannot be reflected in the extracted data. The other option is to include only chain 2 and select a single cognacy judgment within that chain (e.g., *cep/ciep/ceh*, *ceh/*(*cep, ciep*), or (*cep, ciep, ceh*)), in which case the information represented by *ciep/*(*cep, ceh*) in chain 1 cannot be reflected in the extracted data. Thus, the limitation of single cognacy judgments inevitably leads to a loss of information.

2.2.1.2 Absence of "Regularity" Category

Next, let us focus on the absence of a "regularity" category in previous data-preparation methods. In Table 5, a linguistic pattern can be observed across several lexical items, in which the consonants *-p*, *-t*, or *-k* following the vowels *a*, *e*, *o*, or *u*, primarily in Hokkaido dialects, correspond to the consonant *-h* in some Sakhalin dialects. This pattern is evident in pairs such as *cep* and *ceh* ('fish'), *apto* and *ahto* ('rain'), *rup* and *ruh* ('ice'), and *raarak* and *raarah* ('smooth'). This correspondence is well known in Ainu linguistics as the *−p/t/k*:*−h/s* correspondence.[3]

In addition, another linguistic pattern can be observed across several lexical items, in which a sequence of two different vowels (i.e., V_1V_2) is reduced to its second vowel (i.e., V_2). This pattern appears, for example, in *ciep* and *cep* ('fish') in Table 5, and in *eramuan*

[3] The consonant *-p/t/k* after the vowel *i* mainly in the Hokkaido Ainu dialects corresponds to the consonant *-s* in some of the Sakhalin Ainu dialects: *sik* (Nos. 1–13, 15, 16, 19–21) and *sis* (Nos 14, 17, 18) in 'eye' of Asai (1974).

and *eraman* (‘know’) in Asai (1974). This pattern can be formalized as the V_1V_2–V_2 criterion.

Because the scope of a lexical item and a lexical form in previous lexicostatistical studies was limited to single words in the basic vocabulary, earlier works (Fukazawa and Ono 2025; Ono and Fukazawa 2025) introduced the concept of “regularity,” which refers to an abstract pattern observed across multiple lexical items. Accordingly, in the following discussion, the *−p/t/k*:*−h/s* correspondence is referred to as *−p/t/k*:*−h/s* regularity, and the V_1V_2–V_2 criterion as V_1V_2–V_2 regularity.[4] The term “rule” is also useful for referring to individual components within a regularity, such as *−p/t/k* rule and *−h/s* rule within *−p/t/k*:*−h/s* regularity, or the V_1V_2 rule and the V_2 rule within V_1V_2–V_2 regularity.

As discussed in this paragraph, the absence of this “regularity” category resulted in the bifurcation of regularity information into two distinct parts: one part was subsumed under the cognacy judgment criteria, rendering it implicitly and thus absent from the extracted data; the other part was treated as noncognate features, appearing redundantly in the extracted data.

Table 7 provides an example of the resulting quantified data obtained from the extracted data in Table 5 under these judgments.

Table 7. Example of quantified data obtained from the extracted data in Table 5. Both V_1V_2–V_2 regularity and *−p/t/k*:*−h/s* regularity are judged as cognate: C(I/E)(P/H): (*ciep, cep, ceh*); A(P/H)TO: (*apto, ahto*); RUYANPE: (*ruyanpe, ruwanpe*); RU(P/H): (*rup, ruh*); RAARA(K/H): (*raarak, raarah*).

	19. fish	76. rain (the rain)					148. ice			167. smooth			
	C(I/E)(P/H)	weni	A(P/H)TO	RUYANPE	atto	sirun	konru	RU(P/H)	tup	rarak	RAARA(K/H)	taarak	testek
1_Yakumo	1	1	0	0	0	0	1	0	0	1	0	0	0
2_Oshamambe	1	1	0	0	0	0	1	0	0	1	0	0	0
3_Horobetsu	1	0	1	0	0	0	1	0	0	1	0	0	0
4_Biratori	1	0	1	0	0	0	1	0	0	1	0	0	0
5_Nukibetsu	1	0	1	0	0	0	1	0	0	1	0	0	0
6_Niikappu	1	0	1	0	0	0	1	0	0	1	0	0	0
7_Samani	1	0	0	1	0	0	1	0	0	1	0	0	0
8_Obihiro	1	0	0	1	0	0	1	0	0	0	0	0	1
9_Kushiro	1	0	0	1	0	0	1	0	0	0	0	0	1
10_Bihoro	1	0	0	1	0	0	1	0	0	0	0	0	1
11_Asahikawa	1	0	0	1	0	0	1	0	0	1	0	0	0
12_Nayoro	1	0	0	1	0	0	1	1	0	0	0	0	1
13_Soya	1	0	0	1	0	0	0	1	0	1	0	0	0
14_Ochiho	1	0	1	0	0	0	0	1	0	0	1	0	0
15_Tarantomari	1	0	0	0	1	0	0	1	0	0	1	0	0
16_Maoka	1	0	1	0	0	0	0	1	0	0	1	0	0
17_Shiraura	1	0	1	0	0	0	0	1	0	0	1	0	0
18_Raichishka	1	0	1	0	0	0	0	1	0	0	1	0	0
19_Nairo	1	0	0	0	1	0	0	0	1	0	0	1	0
20_North Kuril	1	0	0	0	0	1	1	0	0	1	0	0	0
21_Chitose	1	0	1	0	0	0	1	0	0	1	0	0	0

[4] This paper also applied the concept of “cognacy judgements” to regularity. For example, *−p/t/k* rule and *−h/s* rule was considered as not corresponding to each other (i.e., “noncognate” in our term) for the purpose of analysis in Asai (1974) but as corresponding to each other (i.e., “cognate” in our term) for the purpose of analysis in Hattori and Chiri (1960) as explained in Ono and Fukazawa (2023, 2024).

Table 8. Example of the resulting quantified data obtained from the extracted data of Table 5. V_1V_2–V_2 regularity is judged as cognate, but *−p/t/k*:*−h/s* regularity as noncognate. C(I/E)P: (*ciep*, *cep*); RUYANPE: (*ruyanpe*, *ruwanpe*).

	19. fish		76. rain (the rain)						148. ice				167. smooth				
	C(I/E)P	ceh	weni	apto	RUYANPE	ahto	atto	sirun	konru	rup	ruh	tup	rarak	raarak	raarah	taarak	testek
1_Yakumo	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
2_Oshamambe	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
3_Horobetsu	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
4_Biratori	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
5_Nukibetsu	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
6_Niikappu	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
7_Samani	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
8_Obihiro	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
9_Kushiro	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
10_Bihoro	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
11_Asahikawa	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
12_Nayoro	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1
13_Soya	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0
14_Ochiho	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
15_Tarantomari	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
16_Maoka	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
17_Shiraura	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
18_Raichishka	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
19_Nairo	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0
20_North Kuril	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0
21_Chitose	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0

Table 9. Example of the resulting quantified data obtained from the extracted data in Table 5. The *−p/t/k*:*−h/s* regularity is judged as cognate, whereas the V_1V_2–V_2 regularity is judged as noncognate. CE(P/H): (*cep, ceh*); A(P/H)TO: (*apto, ahto*); RUYANPE: (*ruyanpe, ruwanpe*); RU(P/H): (*rup, ruh*); RAARA(K/H): (*raarak, raarah*).

	19. fish		76. rain (the rain)					148. ice			167. smooth			
	ciep	CE(P/H)	weni	A(P/H)TO	RUYANPE	atto	sirun	konru	RU(P/H)	tup	rarak	RAARA(K/H)	taarak	testek
1_Yakumo	1	0	1	0	0	0	0	1	0	0	1	0	0	0
2_Oshamambe	1	0	1	0	0	0	0	1	0	0	1	0	0	0
3_Horobetsu	1	1	0	1	0	0	0	1	0	0	1	0	0	0
4_Biratori	0	1	0	1	0	0	0	1	0	0	1	0	0	0
5_Nukibetsu	0	1	0	1	0	0	0	1	0	0	1	0	0	0
6_Niikappu	0	1	0	1	0	0	0	1	0	0	1	0	0	0
7_Samani	0	1	0	0	1	0	0	1	0	0	1	0	0	0
8_Obihiro	0	1	0	0	1	0	0	1	0	0	0	0	0	1
9_Kushiro	0	1	0	0	1	0	0	1	0	0	0	0	0	1
10_Bihoro	0	1	0	0	1	0	0	1	0	0	0	0	0	1
11_Asahikawa	0	1	0	0	1	0	0	1	0	0	1	0	0	0
12_Nayoro	0	1	0	0	1	0	0	1	1	0	0	0	0	1
13_Soya	0	1	0	0	1	0	0	0	1	0	1	0	0	0
14_Ochiho	0	1	0	1	0	0	0	0	1	0	0	1	0	0
15_Tarantomari	0	1	0	0	0	1	0	0	1	0	0	1	0	0
16_Maoka	0	1	0	1	0	0	0	0	1	0	0	1	0	0
17_Shiraura	0	1	0	1	0	0	0	0	1	0	0	1	0	0
18_Raichishka	0	1	0	1	0	0	0	0	1	0	0	1	0	0
19_Nairo	0	1	0	0	0	1	0	0	0	1	0	0	1	0
20_North Kuril	0	1	0	0	0	0	1	1	0	0	1	0	0	0
21_Chitose	1	1	0	1	0	0	0	1	0	0	1	0	0	0

Both the V_1V_2–V_2 regularity—typically represented by *ciep* (Nos. 1–3) and *cep* (Nos. 3–13, 15, 19–21) in Table 5—and the *−p/t/k*:*−h/s* regularity—typically represented by *cep* (Nos. 4–13, 15, 19–21) and *ceh* (Nos. 14, 16–18) in Table 5—disappear in Table 7. As a

result, the extracted data for ‘fish’ are reduced to a single summarized lexical form, namely C(I/E)(P/H).

Second, if the V_1V_2–V_2 regularity is treated as cognate, whereas the *−p/t/k*:*−h/s* regularity is treated as noncognate, the classification becomes *ceh*/(*cep*, *ciep*) in ‘fish’. Here, the V_1V_2–V_2 regularity is obscured, whereas the *−p/t/k*:*−h/s* regularity manifests redundantly across other lexical items in the extracted data.

Table 8 provides an example of the resulting quantified data obtained from the extracted data in Table 5 under these judgments. In Table 8, the V_1V_2–V_2 regularity has disappeared, and the extracted data for ‘fish’ consist of one summarized lexical form (i.e., C(I/E)H) and one unsummarized lexical form (i.e., *ceh*).

Third, if the V_1V_2–V_2 regularity is selected as noncognate and the *−p/t/k*:*−h/s* regularity as cognate, the classification becomes *ciep*/(*cep*, *ceh*). Table 9 is the example of the resulting quantified data obtained from the extracted data of Table 5 in these judgments. In Table 9, the *−p/t/k*:*−h/s* regularity has disappeared, and the extracted data for ‘fish’ consist of one summarized lexical form (i.e., CE(P/H)) and one unsummarized lexical form (i.e., *ciep*).

Finally, if researchers select both criteria as noncognate, the classification is *cep*/*ciep*/*ceh* in ‘fish’. Table 6 is the example of the resulting quantified data obtained from the extracted data of Table 5 in these judgments. Consequently, the *−h/s* rule in *−p/t/k*:*−h/s* correspondence is repeatedly reflected in the extracted information in the gray cell on Table 6 (i.e,. *ceh*, *ahto*, *ruh*, and *raarah* [Nos. 14, 16−18]).

Thus, owing to the absence of a “regularity” category, Tables 7–9 result in the loss of information concerning the V_1V_2–V_2 regularity, the *−p/t/k*:*−h/s* regularity, or both.

2.2.1.3 Artificially Inflated Patterns in Quantified Data

Next, let us focus on the problem of repeatedly observed patterns in details. Regardless of the specific algorithm used, quantification methods are generally grounded in pattern recognition of the extracted data. As such, the redundancy of patterns plays a critical role.

For example, *−h/s* rule in *−p/t/k*:*−h/s* regularity is not only in *ceh*, *ahto*, *ruh*, and *raarah* (Nos. 14, 16−18) on Tables 6 and 8 but in other numerous lexical items[5] used in Asai (1974), resulting in the recurrent pattern in the quantified data due to the absence of “regularity” category.

Actually, the absence of the “regularity” category has been the primary factor leading to the classification of “Central Sakhalin” (Nos. 14, 16–18) in Asai’s (1974) cluster analysis, as shown in Figure 1.

[5] For example, *rap* (Nos. 1−13, 15, 20, 21) and *rah* (Nos. 14, 16−18) in ‘36. feather’; *imak* (Nos. 7−10, 13, 15, 19, 20) and *imah* (Nos. 14, 16−18) in ‘43. tooth’; *apkas* (Nos. 1−13, 20, 21) and *ahkas* (Nos. 14, 16−18) in ‘65. walk’; *tek* (Nos. 1−13, 15, 19−21) and *teh* (Nos. 14, 16−18) in ‘66. come’; *cup* (Nos. 1−13, 19−21) and *cuh* (Nos. 14, 16−18) in ‘72. sun’.

Table 10. Similarity matrix for 21 Ainu dialects across 110 lexical items in Asai (1974), based on the "relation index" (Asai 1974: 61−62) and recalculated by Ono and Fukazawa (2023: 106; Table 5). Column numbers correspond to the place number in row, respectively. Values in each cell indicate the degree of similarity between pairs of dialects. As indicated by the single rectangle, the Yakumo dialect (No. 1 in Figure 1) shares at least one cognate lexical form with the Oshamambe dialect (No. 2 in Figure 1) in 104 of the 110 lexical items used in Asai (1974).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1_Yakumo	110	104	102	97	97	96	83	90	90	85	95	90	80	39	53	42	38	42	42	55	95
2_Oshamambe	104	110	100	95	94	93	77	85	84	78	92	87	77	36	49	39	35	40	40	50	94
3_Horobetsu	102	100	110	104	99	101	83	92	91	86	97	94	78	40	56	43	39	43	44	59	99
4_Biratori	97	95	104	110	103	105	79	87	85	80	91	89	74	40	58	45	40	44	46	58	103
5_Nukibetsu	97	94	99	103	110	102	76	82	82	75	88	85	75	36	51	39	34	39	39	54	99
6_Niikappu	96	93	101	105	102	110	77	84	84	78	89	84	73	36	52	39	35	39	40	54	99
7_Samani	83	77	83	79	76	77	110	94	91	83	84	81	74	29	48	32	29	30	34	51	78
8_Obihiro	90	85	92	87	82	84	94	110	106	100	95	93	81	35	55	41	36	37	43	59	86
9_Kushiro	90	84	91	85	82	84	91	106	110	101	98	98	84	35	55	40	36	39	42	63	86
10_Bihoro	85	78	86	80	75	78	83	100	101	110	92	91	78	33	53	40	34	37	40	58	79
11_Asahikawa	95	92	97	91	88	89	84	95	98	92	110	101	82	41	59	45	38	43	43	61	93
12_Nayoro	90	87	94	89	85	84	81	93	98	91	101	110	85	41	60	48	39	45	46	61	87
13_Soya	80	77	78	74	75	73	74	81	84	78	82	85	110	45	66	55	46	49	52	52	74
14_Ochiho	39	36	40	40	36	36	29	35	35	33	41	41	45	110	66	92	90	84	60	29	37
15_Tarantomari	53	49	56	58	51	52	48	55	55	53	59	60	66	66	110	78	67	67	72	45	55
16_Maoka	42	39	43	45	39	39	32	41	40	40	45	48	55	92	78	110	92	88	66	31	43
17_Shiraura	38	35	39	40	34	35	29	36	36	34	38	39	46	90	67	92	110	93	65	28	36
18_Raichishka	42	40	43	44	39	39	30	37	39	37	43	45	49	84	67	88	93	110	60	35	40
19_Nairo	42	40	44	46	39	40	34	43	42	40	43	46	52	60	72	66	65	60	110	37	41
20_North_Kuril	55	50	59	58	54	54	51	59	63	58	61	61	52	29	45	31	28	35	37	110	53
21_Chitose	95	94	99	103	99	99	78	86	86	79	93	87	74	37	55	43	36	40	41	53	110

Table 10 provides an example of visualization applied to the quantified data used in Asai (1974), in which similarity is calculated using the definition of the "relation index" (Asai 1974: 61−62).[6]

We observe that the values among the Ochiho, Maoka, Shiraura, and Raichishka dialects (Nos. 14, 16–18), highlighted in the shaded cells of Table 10, show relatively higher similarity within the Sakhalin Ainu dialects enclosed by the large rectangle in the table. This inflated similarity can be reduced by summarizing the repeated patterns into a single *−p/t/k*:*−h/s* regularity when a "regularity" category is introduced.

2.2.2 A Solution to Previous Methodological Problems in Data Preparations

Given the limitations of existing data-preparation methods, our proposed approach incorporates "multiple cognacy judgments" and a new "regularity" category. For specific details on the extraction method, see Ono and Fukazawa (2025)

Table 11 provides an example of the resulting quantified data obtained from the extracted data in Table 5 using our proposed data-preparation methods. By allowing multiple cognacy judgments, we can avoid the loss of information caused by restricting data-preparation methods in a single cognacy judgment.

[6] Note that *−p/t/k*:*−h/s* regularity was treated as noncognate in Asai (1974).

Table 11. Example of the resulting quantified data obtained from the extracted data through our new data-preparation methods applied to Table 5.

	19. fish	76. rain (the rain)					148. ice			167. smooth				*−p/t/k*:*−h/s* correspondence		V_1V_2–V_2 regularity	
	C(I/E)(P/H)	weni	A(P/H)TO	RUYANPE	atto	sirun	konru	RU(P/H)	tup	rarak	RAARA(K/H)	taarak	testek	−p/t/k	−h/s	V_1V_2	V_2
1 Yakumo	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0
2 Oshamambe	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0
3 Horobetsu	1	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	1
4 Biratori	1	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1
5 Nukibetsu	1	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1
6 Niikappu	1	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1
7 Samani	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1
8 Obihiro	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1
9 Kushiro	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1
10 Bihoro	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1
11 Asahikawa	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1
12 Nayoro	1	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	1
13 Soya	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	1
14 Ochiho	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0.5	0.5
15 Tarantomari	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0.5	0.5
16 Maoka	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0.5	0.5
17 Shiraura	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0.5	0.5
18 Raichishka	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0.5	0.5
19 Nairo	1	0	0	0	1	0	0	0	1	0	0	1	0	1	0	0.5	0.5
20 North Kuril	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0.5	0.5
21 Chitose	1	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1

Table 12. All combinations of cognacy judgments for the V_1V_2–V_2 regularity and the *−p/t/k*:*−h/s* regularity, together with the corresponding disadvantages associated with each combination from the perspective of previous data-preparation methods in lexicostatistics. All combinations exhibit some form of disadvantage.

	V_1V_2-V_2 regularity	−p/t/k:−h/s regularity	disadvantage in V_1V_2–V_2 regularity	disadvantage in −p/t/k:−h/s regularity
Table 6	noncognate	noncognate	repeatedly observed in extracted data	repeatedly observed in extracted data
Table 7	cognate	cognate	NOT in extracted data	NOT in extracted data
Table 8	cognate	noncognate	NOT in extracted data	repeatedly observed in extracted data
Table 9	noncognate	cognate	repeatedly observed in extracted data	NOT in extracted data

Moreover, introducing the “regularity” category serves a dual purpose: it not only eliminates the bias caused by artificially inflated patterns but also resolves the issue where regularity information, previously treated merely as cognate, was not reflected in the extracted lexical items. As a result, the recurrent pattern observed in Tables 6 and 8 no longer appears in Table 11 as lexical-item data but appears as regularity data.

Table 12 summarizes Tables 6–9 with respect to two points: first, the cognacy judgments applied to the V_1V_2–V_2 regularity and the *−p/t/k*:*−h/s* regularity; and second, the resulting disadvantages affecting the representation of these regularities in the extracted data. As shown in Table 12, extracted data produced by previous data-preparation methods necessarily exhibit disadvantages with respect to regularity representation. By contrast, Table 11, which is obtained through our proposed data-preparation methods, does not exhibit disadvantages for either the V_1V_2–V_2 regularity or the *−p/t/k*:*−h/s* regularity. Both regularities are reflected in the extracted data, but they do not appear in an artificially repetitive manner. This outcome demonstrates that our data-preparation methods are superior to previous approaches in maximizing the information obtainable from lexicostatistical documents and in debiasing artificially repeated patterns in the extracted data.

Specifically, our data preparations first exhaustively extracted regularities from the lexicostatistical documents. Subsequently, the remaining information parts of lexical

items not encompassed by these regularities were recorded as lexical-item data. We used two distinct datasets for quantification: one consisting solely of lexical-item data, and the other combining lexical-item data with regularities. For detailed specifications of the constructed datasets, see Fukazawa and Ono (2026).

2.2.3 A Conceptual Framework of New Data-Preparation Methods

We will formalize the conceptual framework of our data-preparation methods to clarify where and how our proposed method outperforms the previous method and to demonstrate the innovative aspects that contribute to methodological developments in this field.

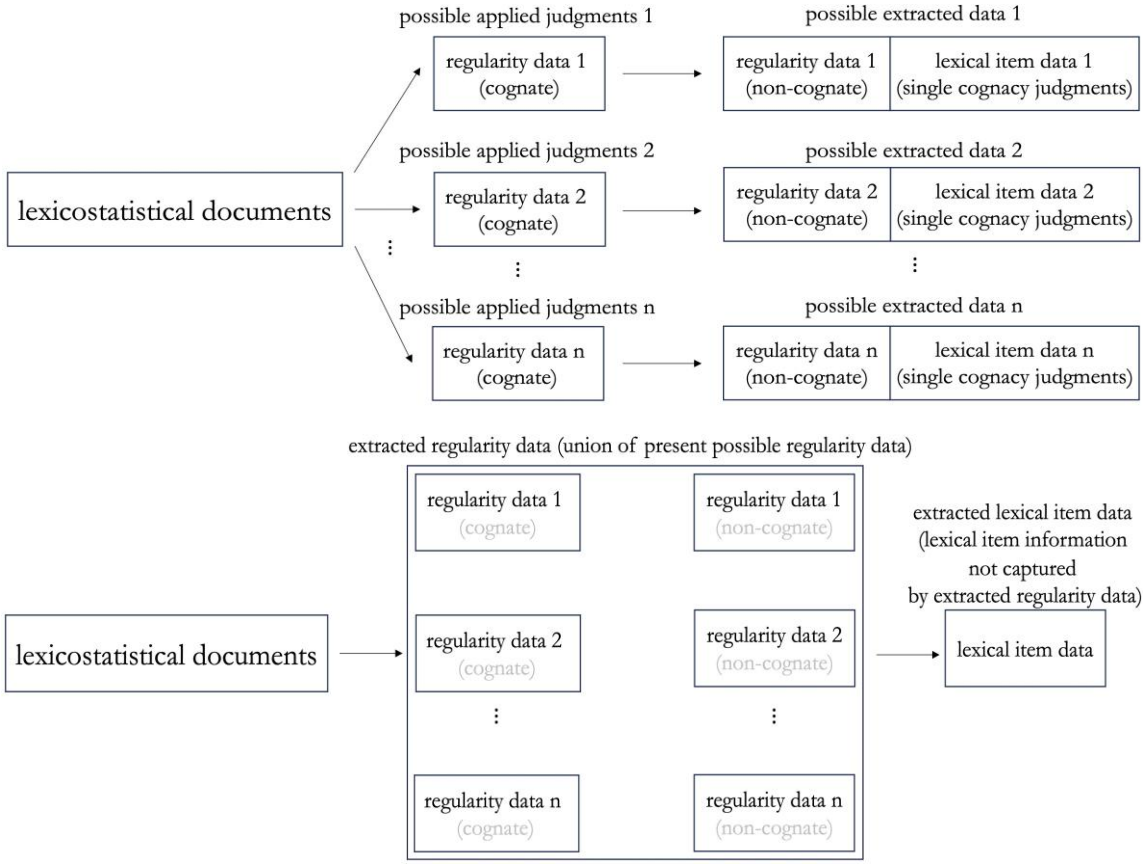

Figure 4. Conceptual framework of our data-preparation methods in lexicostatistics.
Top: previous data-preparation method; Bottom: our proposed data-preparation methods.

As shown at the top of Figure 4, in previous lexicostatistical data-preparation methods, the extracted data depend on specific judgments applied to the lexicostatistical documents (i.e., regularity data that researchers judged as cognate). Furthermore, these data are limited to a single cognacy judgment for each lexical item.

For example, Tables 6–9 can be interpreted as possible extracted data 1–4,

respectively, at the top of Figure 4. Possible extracted data 1 (i.e., Table 6) consists of two components: noncognate V_1V_2–V_2 regularity and *−p/t/k*:*−h/s* regularity, and lexical-item data 1, such as the remaining portions of lexical forms represented in Table 6.

As shown at the bottom of Figure 4, our proposed data-preparation methods address the limitation of previous methods that focus mainly on lexical items rather than regularity information. Thus, our methods first concentrate on maximizing the extracted regularity data, treating them as a union of possible regularities, regardless of whether they are considered cognate or noncognate. Moreover, the lexical item data are redefined as secondary extracted data, representing lexical information not captured by our extracted regularity data (cf. Fukazawa and Ono [2026]).

For instance, Table 11 exemplifies the extracted data produced by our data-preparation methods. The extracted regularity data are not limited to the V_1V_2–V_2 regularity and the *−p/t/k*:*−h/s* regularity, but also include other possible regularities, such as the *y:w* regularity (e.g., *ruyanpe* and *ruwanpe* in ‘rain’) and the *rV:tV* regularity (e.g., *rup* and *tup* in ‘ice’, and *raarak* and *taarak* in ‘smooth’), as discussed in Fukazawa and Ono (2026).

Notably, the extracted regularity data can be interpreted as information relevant to the historical linguistics of the Ainu language, because these data consist of several linguistic correspondences (e.g., *−p/t/k*:*−h/s* correspondence) that are generally understood to reflect historical divergence among languages or dialects.

By contrast, linguistic information not captured by the extracted regularity data is recorded as lexical-item data in the second step of our data-preparation methods. Examples include C(I/E)(P/H) for ‘fish’, A(P/H)TO and RUYANPE for ‘rain’, RU(P/H) for ‘ice’, and RAARA(K/H) for ‘smooth’. We note that the extracted lexical-item data tend to summarize lexical forms to their roots as a consequence of this procedure.

In this way, our data-preparation methods conceptually separate historical information about languages or dialects, which is represented as regularity data, from geolinguistic information including language contact phenomena such as borrowing, which is represented as lexical-item data. A dataset that integrates both regularity data and lexical-item data can therefore provide unified linguistic information relevant to historical linguistics, including geolinguistics and comparative linguistics, while maximizing the extracted information and avoiding bias introduced by artificially repeated patterns in previous data-preparation methods.

2.2.4 Terminologies in Our Data-Preparation Methods

In this paragraph, the terminologies in our data preparation are summarized using the example data shown in Table 13. The basic extraction principles were explained in detail in Ono and Fukazawa (2025).

Table 13. Example of lexical-item data and regularity data extracted by our data-preparation methods. Lexical-item data are extracted from the basic word of 'ice', and regularity data are extracted from the lexical items related to *−p/t/k*:*−h/s* phonological correspondence, such as 'fish' (*cep*, *ciep*, *ceh*) and 'ice' (*rup*, *ruh*). The presence of lexical forms (i.e., *konru*, RU(P/H), and *tup*) and rules (i.e., *−p/t/*k and *−h/s*) is coded as 1, and the absence as 0 in the corresponding dialect in row, respectively.

		lexical item 1: 'ice' (L1)			regularity 1: −p/t/k:−h/s correspondence (R1)	
		konru (L1_1)	RU(P/H) (L1_2)	tup (L1_3)	−p/t/k (R1_1)	−h/s (R1_2)
1_Yakumo	D1	1	0	0	1	0
2_Oshamambe	D2	1	0	0	1	0
3_Horobetsu	D3	1	0	0	1	0
4_Biratori	D4	1	0	0	1	0
5_Nukibetsu	D5	1	0	0	1	0
6_Niikappu	D6	1	0	0	1	0
7_Samani	D7	1	0	0	1	0
8_Obihiro	D8	1	0	0	1	0
9_Kushiro	D9	1	0	0	1	0
10_Bihoro	D10	1	0	0	1	0
11_Asahikawa	D11	1	0	0	1	0
12_Nayoro	D12	1	1	0	1	0
13_Soya	D13	0	1	0	1	0
14_Ochiho	D14	0	1	0	0	1
15_Tarantomari	D15	0	1	0	1	0
16_Maoka	D16	0	1	0	0	1
17_Shiraura	D17	0	1	0	0	1
18_Raichishka	D18	0	1	0	0	1
19_Nairo	D19	0	0	1	1	0
20_North Kuril	D20	1	0	0	1	0
21_Chitose	D21	1	0	0	1	0

We define a lexical item as comprised of several lexical forms and regularity as consisting of several rules. For example, the extracted lexical item of 'ice' consists of three lexical forms (i.e., *konru*, RU(P/H), and *tup*), and the extracted regularity of the *−p/t/k*:*−h/s* phonological correspondence consists of two rules (i.e., *−p/t/k* and *−h/s*). Note that our specific symbolization methods applied to Asai's lexicostatistical documents are described in Fukazawa and Ono (2026).

As discussed in following subsection, homogeneity analysis will assign the numbers, which can capture some structures in the data from specific numerical criteria, not only to each lexical form and each rule (e.g., L1_1, L1_2, L1_3, R1_1, R1_2) but also to each dialect (e.g., D1, D2,...,D21), using the extracted information of lexicostatistical documents as in Figure 2. Moreover, the additional information about each lexical item or regularity is assigned as numbers (e.g., L1 and R1).

2.3 Quantification Methods and Quantified Data

This subsection primarily explains our proposed quantification methods, using Table 14 (Table 6) as an illustrative example.

2.3.1 Methodological Problems in Quantification Methods

Table 14 provides a typical example of quantified data derived from Table 5, in which the presence of a lexical form is coded as 1 and its absence as 0.

The methodological problems associated with previous quantification methods can be summarized in four points: binary scoring, unidimensional quantification, quantified scores limited to human recognition, and the low explainability of dialect scores.

Table 14. The quantified data of Table 5.

	19. fish			76. rain (the rain)						148. ice				167. smooth				
	ciep	cep	ceh	weni	apto	RUYANPE	ahto	atto	sirun	konru	rup	ruh	tup	rarak	raarak	raarah	taarak	testek
1_Yakumo	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
2_Oshamambe	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
3_Horobetsu	1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
4_Biratori	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
5_Nukibetsu	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
6_Niikappu	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0
7_Samani	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
8_Obihiro	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
9_Kushiro	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
10_Bihoro	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
11_Asahikawa	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
12_Nayoro	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1
13_Soya	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0
14_Ochiho	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
15_Tarantomari	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
16_Maoka	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
17_Shiraura	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
18_Raichishka	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
19_Nairo	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0
20_North Kuril	0	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0
21_Chitose	1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0

For example, in Table 14, binary scores (i.e., 1 or 0) are assigned to the presence of the lexical forms *ciep*, *cep*, and *ceh* in ‘fish’. However, there is a gradable degree of similarity among these lexical forms, such as *ciep* < *cep* < *ceh* (i.e., an order relation). Because such an order relation among three elements cannot be represented by binary scores, our proposed quantification methods assign continuous values to each lexical form within a lexical item, or to each rule within a regularity, for example, −0.2 for *ciep*, 0.2 for *cep*, and 0.6 for *ceh*.

Second, consider a case in which continuous scores are assigned to *apto* as −0.2, *ahto* as 0.2, and *atto* as 0.6 in ‘rain (the rain)’, yielding the order *apto* < *ahto* < *atto*. From the perspective of Ainu linguistics, however, an alternative ordering is also plausible, such as *apto* < *atto* < *ahto*. In fact, many previous quantification methods were restricted to unidimensional scoring schemes that forced a single linear ordering, such as *apto* < *ahto* < *atto* or *apto* < *atto* < *ahto*.

By contrast, multidimensional scoring allows *apto* < *ahto* < *atto* to be represented along a first dimension, and *apto* < *atto* < *ahto* along a second dimension. In this way, multidimensional quantification can capture more complex historical information

encoded in lexical forms.

Third, Table 14 exhibits several geographically structured distributions, including the Hokkaido dialect group in *konru* (Nos. 1–12, 20, 21), the Saru–Chitose type in *apto* (Nos. 3–6, 21), and the northeastern Hokkaido dialects in RUYANPE (Nos. 7–13). Previous quantification methods relied heavily on human pattern-recognition abilities, but such linguistic patterns must be identified across 353 lexical forms in 87 lexical items and 49 rules in 22 regularities in our dataset.

Moreover, these geographically distributed dialect patterns sometimes exhibit apparent contradictions. For example, when examining the common relationship between the dialectal distributions of *cep* (Nos. 3–13, 15, 19–21) and *apto* (Nos. 3–6, 21), the Yakumo and Oshamambe dialects (Nos. 1–2) are excluded. By contrast, when examining the common relationship between the distributions of *rarak* (Nos. 1–7, 11, 13, 20, 21) and *apto* (Nos. 3–6, 21), the Yakumo and Oshamambe dialects (Nos. 1–2) are included. This indicates that the dialectal pattern associated with *apto* (Nos. 3–6, 21) occupies a contradictory position, either including or excluding the Yakumo and Oshamambe dialects. Consequently, our proposed quantification methods, implemented as computer algorithm, must be capable of automatically extracting multiple dialectal patterns, even when these patterns are potentially contradictory, from the large volume of information contained in the extracted data.

Finally, consider the similarity matrix shown in Table 10, which was calculated by counting shared lexical forms between each pair of Ainu dialects using the "relation index" defined by Asai (1974: 61–62). When further visualization techniques, such as the cluster analysis employed by Asai (1974), are applied to Table 10, dialect classification can be represented as a hierarchical tree structure (i.e., a dendrogram; see Figure 2 in the supplementary materials). However, Table 10 itself does not retain information about which lexical items or regularities contribute to higher or lower similarity values between dialect pairs.

Because a central objective of our proposed lexicostatistical methodologies is to shift the focus from classification back to historical relationships among languages or dialects, improving the explainability of dialect scores and positions is a critical requirement for our new quantification methods.

2.3.2 A Solution to Previous Methodological Problems in Quantification Methods

To address these issues, we employ homogeneity analysis (Gifi 1990) as the core quantification method in our lexicostatistical framework, assigning numerical scores to the extracted data. The computational procedures of homogeneity analysis are summarized in Figure 5, and a complete, worked example of homogeneity analysis applied to Table 13 is provided in Section 3 on the supplementary materials.

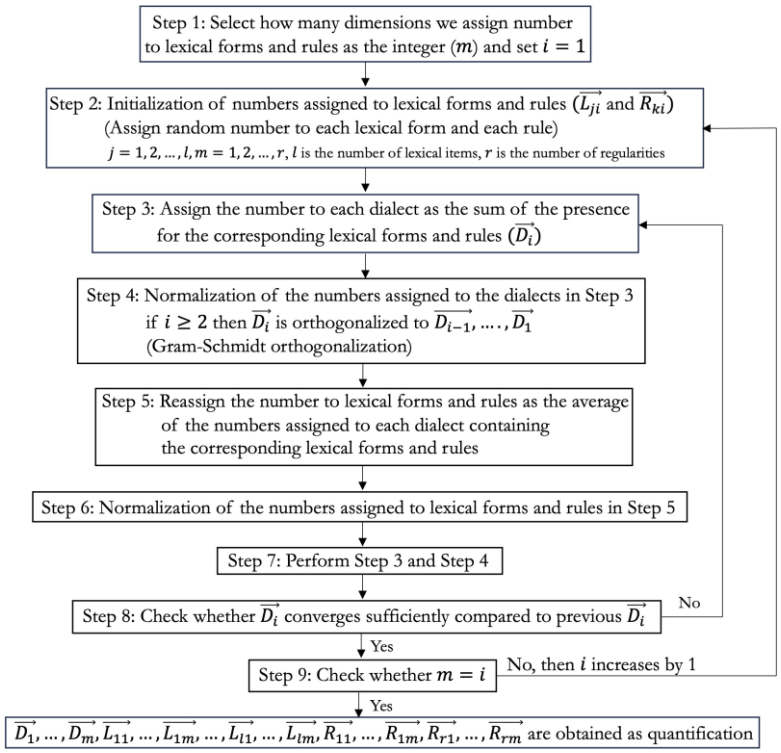

Figure 5. Flowchart of homogeneity analysis applied to the dataset, combining both lexical-item and regularity data from the perspective of the method of reciprocal averages. Note that the flowchart of homogeneity analysis applied to only lexical-item data corresponds to the flowchart without the information of the vector of R in this figure.

Let us examine how homogeneity analysis addresses the problems inherent in previous quantification methods, using Table 13 as an illustrative example. First, continuous numerical values are assigned to *konru* as L1_1, RU(P/H) as L1_2, *tup* as L1_3, the *−p/t/k* rule as R1_1, and the *−h/s* rule as R1_2 (with lexical item 1 denoted as L1 and regularity 1 as R1). In this way, homogeneity analysis achieves continuous quantification.

Second, when researchers set the value of *m* as 3 in Step 1 of Figure 5, distinct numerical values are assigned to L1_1, L1_2, L1_3, R1_1, and R1_2 in each of the three dimensions. As a result, homogeneity analysis realizes multidimensional quantification.

Third, homogeneity analysis enables the simultaneous quantification of a large number of lexical forms and rules through computer programs implemented in the R language (R Core Team 2022; de Leeuw and Mair 2009). Moreover, the Gram–Schmidt orthogonalization procedure (Gram 1883; Schmidt 1907) employed in homogeneity analysis (Step 4 in Figure 5) allows researchers to assign continuous, multidimensional scores in a particularly informative manner. Specifically, the scores assigned to the third dimension capture residual linguistic information that is not represented in the scores of the first dimension (e.g., primarily reflecting differences between North Kuril dialects and

others) or the second dimension (e.g., primarily reflecting differences between Sakhalin Ainu dialects and others).

In this way, homogeneity analysis overcomes the limitations of human pattern recognition, which tends to focus on more salient contrasts in the data while overlooking subtler patterns that do not align with dominant differences. Further details are provided in the supplementary materials.

Finally, as shown in Step 3 of Figure 5, continuous numerical values are assigned to each dialect as D1–D21 in homogeneity analysis. Each dialect score approximates the sum of the values assigned to the lexical forms or rules present in that dialect.

For example, the Yakumo dialect (No. 1) in Table 13 includes the lexical form *konru* for 'ice' and the *−p/k/t* rule within the *−p/k/t*:*−h/s* regularity. Accordingly, the value of D1 is calculated as the sum of the value assigned to *konru* in Step 2 (L1_1) and the value assigned to the *−p/k/t* rule in Step 2 (R1_1). Consequently, when homogeneity analysis is applied separately to lexical-item data and to a dataset combining lexical-item and regularity data, researchers can identify which lexical forms or rules are responsible for shifts in the relative positions of dialects between the two configurations. This is achieved by comparing the values assigned to lexical forms in the lexical-item dataset with those assigned to lexical forms and rules in the combined dataset, as discussed in Section 3.[7]

Taken together, these properties indicate that homogeneity analysis is a promising solution to the four methodological problems identified in previous quantification methods in lexicostatistics. In this paper, our lexicostatistical framework adopts the revised homogeneity analysis proposed by Ono and Fukazawa (2025), which is hereafter referred to simply as "homogeneity analysis."

2.4 Visualization and Obtainable Linguistic Information

This subsection explains our proposed visualization methods and their advantages. We begin with Table 4, which visualizes dialectal patterns by sorting lexical forms. Sorting represents a basic visualization technique that renders linguistic groupings in the data visible.

However, sorting has several limitations. First, although sorting can reveal multiple groupings of dialects—such as the Saru–Chitose type (Nos. 4–6, 21) in HUNNA and MAK, the Saru–Chitose–Sakhalin type (Nos. 4–6, 14–19, 21) in APTO, the northeastern Hokkaido dialects (Nos. 8–12) in EKACI, or the Hokkaido dialects other than the Saru–Chitose type (Nos. 1–3, 7–13) in *nen* and NEKONA—the relative strength of each classification remains unclear. Moreover, the relative positions of these dialects are difficult to grasp intuitively through sorting alone.

[7] Another quantification methods (e.g., correspondence analysis [Benzécri et coll. 1973]) cannot achieve such high explainability to dialect scores.

For example, Table 4 does not indicate which classification is strongest among the Saru–Chitose type, the Saru–Chitose–Sakhalin type, the northeastern Hokkaido dialects, or the Hokkaido dialects other than the Saru–Chitose type, nor does it clearly show which dialect lies closest to the Soya dialect (No. 13).

To address such issues, statisticians have proposed various alternatives, including Hayashi's quantification methods (Hayashi 1950), correspondence analysis (Benzécri et coll. 1973), dual scaling (Nishisato 1980), and homogeneity analysis (Gifi 1990). Nevertheless, a fundamental limitation has persisted: visualization results have traditionally been constrained to two-dimensional (2D) representations, because research articles were published in print prior to advances in computing and internet technologies. As a result, multidimensional scores obtained through quantification methods were necessarily projected onto a 2D plane, leading to the loss of linguistic information encoded in higher dimensions.

Recent advances in web-based computational technologies, however, have transformed data visualization practices. In particular, Hyper Text Markup Language 5 (HTML5; standardized as HTML Living Standard from 2021) enables the visualization of 3D linguistic information within standard web browsers such as Google Chrome and Microsoft Edge.

Furthermore, HTML5 allows 3D linguistic information to be visualized not only statically but also dynamically. As demonstrated in the HTML5 files provided in the supplementary materials, users can interact with the visualization through rotation, zooming, and scaling within a 3D space, as well as access detailed information at specific dialect points, such as values along the first, second, and third dimensions.

As a result, the scope of obtainable information—previously constrained by information loss inherent in static 2D maps—has expanded substantially. Whereas earlier lexicostatistical visualization methods tended to restrict the field to mere "classification," often compounded by biases in the extracted and quantified data, the implementation of 3D visualization enables a renewed focus on historical linguistic information. This shift is made possible because 3D visualization can exploit the richer structure of multidimensional quantified scores, and because our data-preparation and quantification methods overcome prior limitations in extracted and quantified data.

Moreover, our lexicostatistical methodologies allow not only the visualization of historical perspectives, such as geolinguistics and comparative linguistics, using datasets that combine lexical-item and regularity data, but also the comparison of results derived from different datasets. Differences between these configurations carry important linguistic implications, supported by the enhanced explainability of dialect scores in homogeneity analysis in terms of specific lexical forms and rules.

In this way, lexicostatistics can shift its focus from classification back to historical inquiry through our proposed methodologies. In the following section, we evaluate

whether these methodologies function as intended when applied to Asai's (1974) lexicostatistical documents.

3. Results

Section 3 presents three-dimensional (3D) visualizations of Ainu dialects derived from a lexical-item dataset and from a dataset combining lexical-item and regularity data, rendered as interactive HTML5 visualizations. We also examine visualizations of lexical items and regularities, focusing on two issues: how the inclusion of regularities alters the configuration of Ainu dialects, and how regularities affect the quantification of lexical items, drawing on the properties of homogeneity analysis discussed in Section 2.

Thus, 3D visualizations of Ainu dialects and quantification scores are presented for lexical items and rules in each dimension. These results are obtained from two sources: the lexical-item data and the combined dataset (comprising both lexical-item and regularity data). Due to space limitations, we provide 2D snapshots of the dialect configurations derived from the original 3D interactive HTML5 graphs, using the *plotly* package (Sievert 2020) in R (R Core Team 2022).

It is available to download the source codes for these graphs from the files NoLS_16_YO_MK_item.txt and NoLS_16_YO_MK_item_regularity.txt, respectively. The original interactive graphs can be accessed by changing the file extension from .txt to .html. Supplementary materials including original 3D graphs can be also downloaded from https://researchmap.jp/ono_yohei/published_papers/51934997?lang=en.

3.1 3D Visualization of Ainu Dialects

Figure 6 shows a 3D visualization of Ainu dialects derived from lexical-item data. The first dimension corresponds to the difference between the Sakhalin and Northern Kuril group (Nos. 14–20) and the Hokkaido Ainu group (Nos. 1–13, 21). The second dimension corresponds to the difference between the Sakhalin dialects and the Northern Kuril dialect, with the Hokkaido group positioned near the average value of zero. The third dimension corresponds to the difference between the northeastern and southwestern Hokkaido groups; it also covers the difference between the Northern and Southern Sakhalin groups.

Next, Figure 7 shows a 3D visualization of Ainu dialects derived from a dataset combining lexical-item and regularity data. The three dimensions roughly correspond to those in Figure 6, respectively.

The key differences between Figures 6 and 7 can be summarized as follows. First, in the third dimension of Figure 6, the Sakhalin dialects (Nos. 14–19) clearly split into a northern group (Nos. 17–19) in the negative direction and a southern group (Nos. 14–16) in the positive direction in the left side of Figure 6. This split is not observed in Figure 7.

Second, in the third dimension of Figure 6, the northern part of the Sakhalin dialects shows negative values similar to those of the southwestern Hokkaido group.

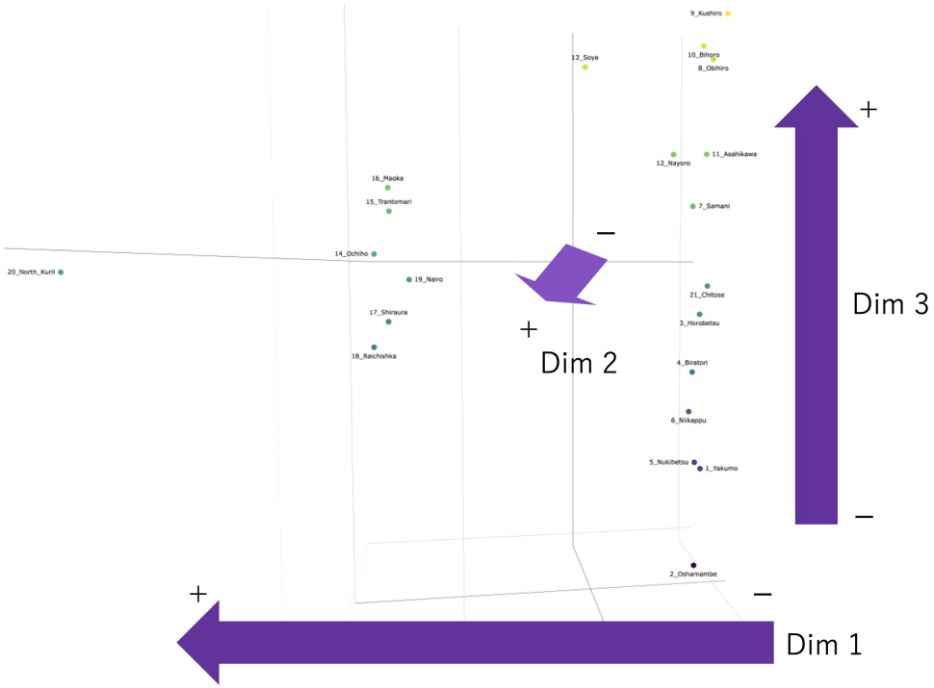

Figure 6. 3D visualization of the quantified scores for Ainu dialects obtained through homogeneity analysis of the lexical-item data prepared in our data-preparation methods.

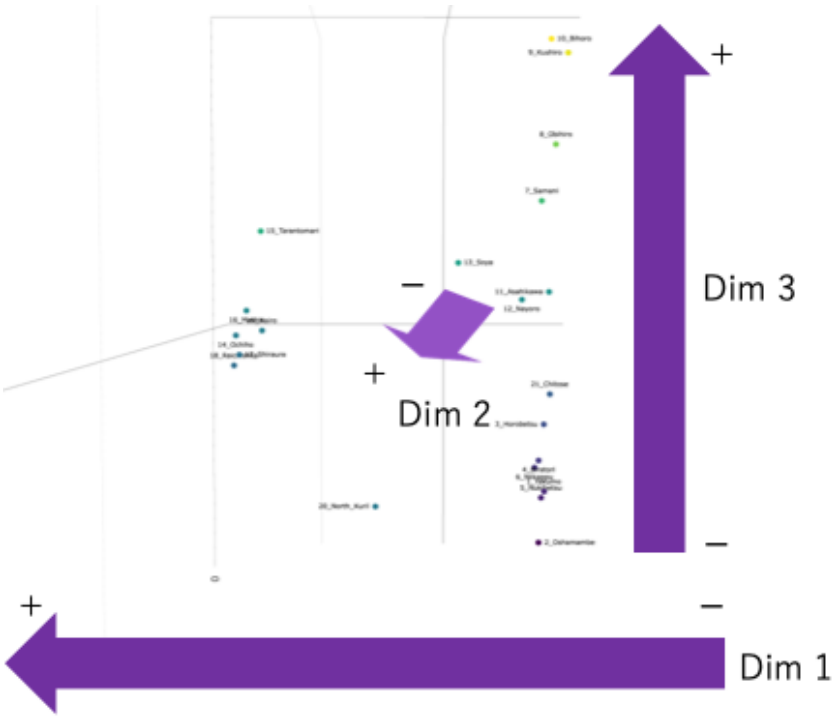

Figure 7. 3D visualization of the quantified scores of Ainu dialects obtained from homogeneity analysis applied to the dataset combining lexical-item and regularity data prepared in our data-preparation methods.

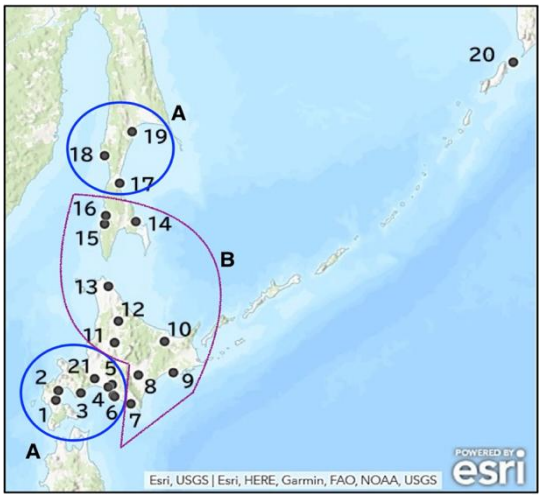

Figure 8. A–B–A distribution in the third dimension on Figure 6. Ainu dialects with negative values on the third dimension in Figure 6 belong to Group A, while those with positive values on the third dimension belong to Group B. Group A is enclosed by blue circles, and Group B by purple circles.

In contrast, the southern part of the Sakhalin dialects shows positive values identical to those of the northeastern Hokkaido group. This pattern, which forms an "A–B–A distribution" in geolinguistics as shown in Figure 8, is present in Figure 6 but absent in Figure 7.

Third, while Figure 6 lacks any dialect positioning around the origin (i.e., no "average" dialect), Figure 7 shows that the Asahikawa, Nayoro, and Soya dialects (Nos. 11–13) occupy positions around the origin (Ono 2019). This suggests that these dialects are clearly pulled toward the origin when regularity data are added to the lexical item data.

3.2 Visualization of Lexical Items and Regularities

Figures 9–11 present the quantification scores for each lexical form and rule, which were shown by the first, second, and third dimensions.

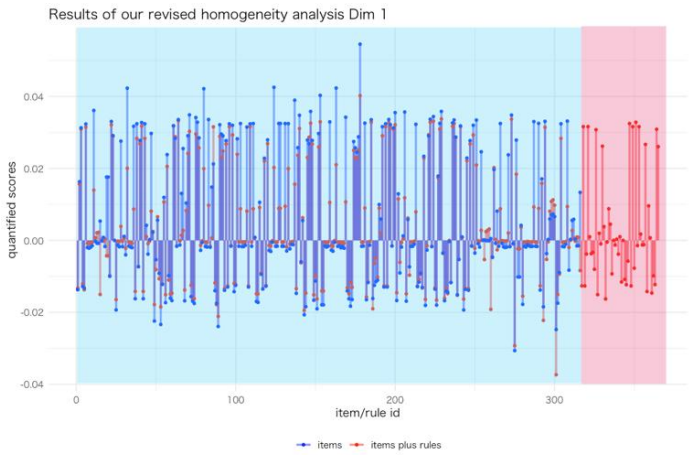

Figure 9. The quantified scores in the first dimension obtained by homogeneity analysis applied to both the lexical-item data and the combined dataset of the lexical-item and regularity data. The lexical-item data are highlighted by blue and the regularity data are highlighted by red.

In this study, the scores derived from lexical-item data (blue graphs) are markedly altered by the addition of regularity data (red graphs). This result stands in contrast to the findings reported in Ono and Fukazawa (2025: 197; Figure 7), which adopted Asai's cognacy judgments and regularities. Moreover, particularly in the third dimension (Figure 11), the

patterns themselves were changed drastically from the scores in lexical-item data only (blue graphs) to those in both lexical-item and regularity data (red graphs)

These results provide statistical evidence that regularity data contain linguistic information distinct from that of lexical-item data. Furthermore, they demonstrate that linguistic information hidden within the lexical-item data can be extracted by incorporating regularity data.

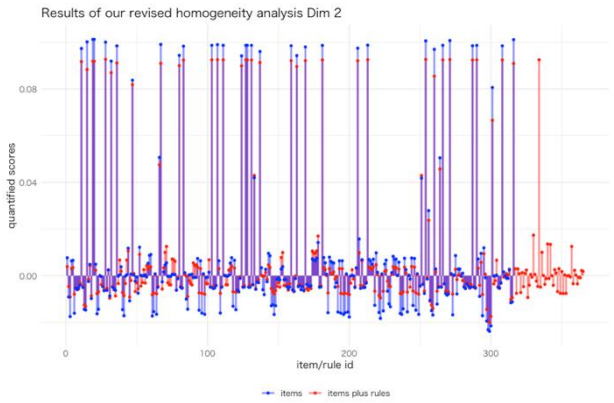

Figure 10. The quantified scores in the second dimension were obtained via homogeneity analysis applied to both the lexical-item data and the combined dataset of lexical-item and regularity data.

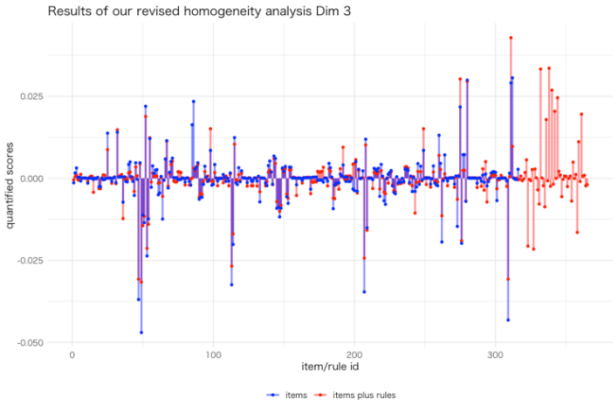

Figure 11. The quantified scores in the third dimension were obtained by applying homogeneity analysis to both the lexical-item data and the combined dataset of lexical-item and regularity data.

4. Discussions

Section 4 depicts the methodological significance of our proposed method from both linguistic and statistical perspectives. Based on the visualization results quantified by homogeneity analysis, we examine whether our proposed method can capture information on historical divergence or language contact in Ainu dialects—features that have previously remained unobserved due to methodological pitfalls in traditional lexicostatistical data preparation.

From a linguistic perspective, our data preparation and quantification methods

revealed significant differences in numerical values assigned to dialects, depending on whether the analysis used only lexical-item data or combined lexical-item data with regularity data. Notably, both approaches provided unique insights for classifying Ainu dialects and interpreting their historical development. Specifically, the results incorporating regularity into lexical items revealed the following: First, the Asahikawa, Nayoro, and Soya dialects formed an intermediate group positioned between the southwestern and northeastern Hokkaido dialect groups.

Second, these dialects were positioned near the origin of the spatial coordinates. In homogeneity analysis, the origin represents the location of a "hypothetical dialect" embodying the average characteristics of the entire dialects. This suggests that the Asahikawa, Nayoro, and Soya dialects possess characteristics close to this average, implying a history of contact and mixing with other dialects.

Third, in the results based solely on lexical-item data, the Sakhalin dialects were clearly divided into northern and southern groups, and fourth, the southern Sakhalin group is dispositioned similar to the northeastern Hokkaido Ainu dialects and the northern Sakhalin group showed marked similarities to the Saru and Chitose dialect group. While the latter corresponds to the "Saru, Chitose, and Sakhalin type" described in Nakagawa (1996) and Nakagawa and Fukazawa (2022), our finding that northern Sakhalin is specifically closer to the Saru and Chitose dialects could have a significant impact on future historical interpretations of the Ainu language.

From a statistical perspective, our clarifications of quantification procedures in Section 2.3 enabled us to distinguish which parts of the visualization are derived from the mathematical properties of the quantification method. For example, eigenvalue formalization in homogeneity analysis can contribute to capturing the first and third dimensions: the former corresponds to the difference with the most significant variation between the Sakhalin and Northern Kuril group (Nos. 14–20) and the Hokkaido Ainu group (Nos. 1–13, 21), and the latter corresponds to the difference with the third largest variation between the northern Sakhalin and southwestern Hokkaido Ainu dialect group and the southern Sakhalin and northeastern Hokkaido dialect group in Figure 6. Moreover, the Gram–Schmidt orthogonalization clearly corresponds to the linguistic insight that the two directions are almost unrelated (Section 3 in supplementary materials).

Clarifications of quantification procedures can demonstrate that some findings in the visualization are newly observed as discoveries. For example, eigenvalue formalism or Gram–Schmidt orthogonalization cannot explain why the third dimension is extracted as an A–B–A distribution in Figure 8.

We consider that previous interdisciplinary collaborative research between linguistics and statistics has not explained the specific calculation procedures in statistical methods or translated their meanings into a linguistic context. It is insufficient because calculation procedures exactly correspond to how the linguistic meaning in the data is transformed

into numbers. This means that it has prevented feedback between disciplines and hindered the resolution of essential problems in their studies, engaged in dialogue with each other about what we can and cannot discuss from the obtained numbers. Thus, our data-preparation and quantification methods are not only innovative in their own right but also an attempt to overcome the fundamental limitations in interdisciplinary research between linguistics and statistics, as exemplified by lexicostatistical documents in Asai (1974).

We propose that our new data-preparation method, together with homogeneity analysis, constitutes a robust approach for uncovering linguistic information of historical change through the integrated analysis of regularities and the refined examination of lexical items.[8]

5. Conclusion

This study elucidated the methodological significance of our novel data-preparation method (Fukazawa and Ono 2025, 2026; Ono and Fukazawa 2025), which uses unique cognacy judgments to generate lexical-item and regularity data, alongside the results and interpretations visualized through homogeneity analysis.

In our previous work, we demonstrated that incorporating inter-dialectal regularities into Asai's (1974) data allowed for the appropriate quantification of regularities Asai had identified (Fukazawa and Ono 2025; Ono and Fukazawa 2025). However, a critical limitation of earlier methods was that recurrent regularities were treated as lexical-item data. Consequently, patterns observed in multiple lexical items—such as the *−p/t/k*:*−h/s* correspondence in Sakhalin and Hokkaido dialects—were artificially amplified during numerical quantification, introducing significant bias into the dialect configuration.

Our new data-preparation method represented a methodological innovation, enabling the visualization of regularity patterns in a bias-free manner. Visualizations derived from both the lexical-item data alone and the combined dataset, including regularity data, indicated the potential to extract historical insights into the Ainu language and dialects that previous research had struggled to capture. Furthermore, the improved homogeneity analysis was proven instrumental in identifying the specific lexical forms and rules deriving dialect distribution. By establishing this methodological foundation, we proposed that lexicostatistics can transcend mere classification and approach the

[8] Moreover, our further studies are still needed to revise the homogeneity analysis, especially to explain why squared differences in loss function of homogeneity analysis can capture linguistic information of historical divergence and language contact. In other realms of physical and psychological science (Luce 1959; Yoshino 1989), the question of which types of function can exist stably within phenomena and capture their essence remains disputed. For example, Nishisato (1984) proposed the method of reciprocal medians rather than the reciprocal averages used in this paper, which does not use squared differences but absolute differences as loss function in the quantification method. Thus, our revised homogeneity analysis is open not only to statistical refinement but also to a mathematical–statistical investigation of the nature of languages or dialects, including their historical divergence and language contact, in future work.

reconstruction of historical change—the fundamental purpose of dialectology.

Acknowledgments

Parts of this paper were presented at the 8th Annual Conference of the Japan Association of Northern Language Studies on September 27, 2025. We are very grateful to the conference organizers and participants, especially Professor Fuyuki Ebata (Niigata University), Professor Norikazu Kogura (Tokyo University of Foreign Studies), Professor Osami Okuda (Sapporo Gakuin University), and Professor Yoshiko Yamada (Muroran Institute of Technology) for their invaluable questions and comments. Finally, we would like to thank the editors and two highly conscientious reviewers for their constructive and invaluable comments; all errors are, of course, our own. JSPS KAKENHI supported this work with grant number JP25K04114.

References

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

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 (2025) Extraction of regularities and geographical patterns from the basic vocabulary of the Ainu language. In Jalaluddin, Nor Hashimah, Hiroyuki Suzuki and Mitsuaki Endō (eds.) *Proceedings of the sixth International Conference of Asian Geolinguistics*, 62–80. doi: https://doi.org/10.5281/zenodo.17204595

Fukazawa, Mika and Yōhei Ono (2025) Kisogoi niyoru hōgen bunrui no shomondai: Asai (1974) no ainugo hōgen deta wo saidaigen katsuyō surutame ni [Challenges of dialect classification in extracting maximum information: from the basic vocabulary of Ainu in Asai (1974)]. *Northern Language Studies*, 15: 141–163.

Fukazawa, Mika and Yōhei Ono (2026) Gokei to kisogoi no fukugōteki na deta no kōchiku: ainugo kisogoi no jōhō no saidaika ni mukete [Building composite data on lexical forms and regularities: extracting maximum information from the basic vocabulary of Ainu dialects]. *Northern Language Studies*, 16: 151–173.

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.

Gram, Jørgen Pedersen (1883) Über die Entwickelung reeller Functionen in Reihen mittelst der Methode der kleinsten Quadrate. *Journal für die Reine und Angewandte Mathematik*, 94: 41–73.

Grätzer, George (1998) *General lattice theory: second edition.* Basel: Birkhäuser Verlag.

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

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

Hayashi, Chikio (1950) On the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 2: 35–47.

Luce, Robert Dancan (1959) On the possible psychophysical laws. *Psychological Review*, 66(2): 81–95.

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

Nakagawa, Hiroshi (1996) Gengochirigaku niyoru ainugo no shiteki kentō [A historical study of the Ainu language through linguistic geography]. *Bulletin of the Hokkaido Ainu Culture Research Center*, 2: 1–18.

Nakagawa, Hiroshi and Mika Fukazawa (2022) Hokkaido Ainu dialects: towards a classification of Ainu dialects. In Anna Bugaeva (ed.) *Handbook of the Ainu language*, 253–328. Berlin: Mouton De Gruyter. doi: https://doi.org/10.1515/9781501502859-009

Nishisato, Shizuhiko (1980) *Analysis of categorical data: dual scaling and its applications.* Toronto: University of Toronto Press.

Nishisato, Shizuhiko (1984) Dual scaling by reciprocal medians. In *Proceedings of the 32nd Scientific Conference of the Italian Statistical Society*, 141–147.

Ono, Yōhei (2019) Observations on "northeastern" Hokkaido Ainu dialects: a statistical perspective. *Northern Language Studies*, 9: 95–122.

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

Ono, Yōhei and Mika Fukazawa (2023) Ainugo shohōgen no gokei no ruiji nikansuru kiso deta no fukugen: ronbun ni kaki kirenakatta kenkyūsha no handan shikō ni 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, Yōhei and Mika Fukazawa (2024) Hikakufukanō datta ainugo hōgenbunrui: tōkeiteki hōgenbunrui wo ruijido no tenkara saikō suru [(Un)comparable classifications of Ainu dialects: reconsidering statistical dialect classification from the similarity judgments]. *Journal of Ainu and Indigenous Studies*, 4: 93–126.

Ono, Yōhei and Mika Fukazawa (2025) Statistical approaches to the quantification of regularity in languages and dialects: an exercise in Ainu dialects of Asai (1974). *Northern Language Studies*, 15: 179–202.

R Core Team (2022) R: A language and environment for statistical computing. R

Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Schmidt, Erhard (1907) Zur Theorie der linearen und nichtlinearen Integralgleichungen I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. *Mathematische Annalen*, 63: 433–476.
Sievert, Carson (2020) *Interactive Web-Based Data Visualization with R, plotly, and shiny.* New York: Chapman and Hall/CRC. URL: https://plotly-r.com.
Swadesh, Morris (1955) Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 24: 121–137.
Torii, Ryūzō (1903) *Chishima Ainu [Kuril Ainu]*. Tokyo: Yoshikawa-Kōbun-Kan.
Yoshino, Ryōzō (1989) On the possible and stable psychophysical laws. *Journal of Mathematical Psychology*, 33(1): 68–90.

Summary

This study proposes a novel data-preparation method in lexicostatistics to potentially uncover historical divergence and language contact in Ainu dialects, addressing the methodological limitations of conventional approaches that redundantly count recurring regularities, thereby introducing statistical bias in previous lexical-item data.

Our method systematically extracts almost all potential regularity data in the first step. Then, it extracts lexical-item data as linguistic information not captured by the regularity data, thereby preventing artificial inflation of specific patterns in previous data. Our revised homogeneity analysis is performed in two datasets: one consisting solely of lexical-item data and another combining lexical items with regularity data.

Quantification results of Ainu dialects are visualized as 3D interactive graphs using HTML5. Visualization results from the dataset, which combines lexical items and regularity data, position the Asahikawa, Nayoro, and Soya dialects near the coordinate origin—representing the "average" characteristics—suggesting historical language contact and mixture in these dialects.

Conversely, the visualization result of lexical-item data revealed a clear north–south division in Sakhalin dialects, with the northern group exhibiting similarity to Saru-Chitose dialects in southwestern Hokkaido Ainu dialects and the southern group showing similarity to northeastern Hokkaido Ainu dialects, demonstrating an A–B–A geolinguistic distribution that has not yet been discovered until our analyses.

These findings demonstrate that our framework can integrate geolinguistic and historical linguistic perspectives: the former aligns with lexical item data in our data-preparation methods, and the latter corresponds to regularity data. Thus, our data-preparation and quantification methods will shift the focus in lexicostatistics from classification back to history in its original interest.

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