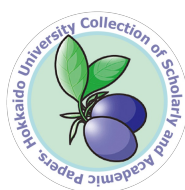




Title	From Classification to History : Examining New Lexicostatistical Methodologies in Using the Data in Ainu Dialects
Author(s)	Ono, Yohei; Fukazawa, Mika
Citation	北方言語研究, 16, 121-150
Issue Date	2026-03-20
DOI	https://doi.org/10.14943/2115.99346
Doc URL	https://hdl.handle.net/2115/99346
Type	departmental bulletin paper
File Information	08_Ono_Fukazawa_Supplementary_Materials_final.pdf, Supplementary Materials



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

Supplementary Materials

1. Study Overview

This paper primarily aims to propose a new data-preparation method in lexicostatistics and to verify its validity, specifically whether historical divergence or language contact in Ainu dialects—which has remained hidden and may not have been captured by current preparation methods—can potentially be revealed by our proposed method.

The problem with preparation methods in lexicostatistics is that information that should have been extracted as a single regularity was instead treated as separate lexical items. Specifically, linguistic patterns across multiple lexical items—such as the $-p/t/k:-h/s$ phonological correspondence in Sakhalin and Hokkaido dialects—were redundantly counted in several lexical items, which artificially inflated the statistical weight of these patterns in quantification and consequently biased the final classification.

In contrast, our proposed data-preparation method is a methodological innovation in collaborative research between linguistics and statistics because it addresses issues by systematically extracting regularity data as the first step. The remaining linguistic information, which cannot be accounted for by the regularity data, is then extracted into lexical item datasets.

This process prevents regularity data from being counted redundantly in the lexical item data. It thereby allows visualization of linguistic patterns in lexicostatistical documents while reducing the bias inherent in current data-preparation methods.

Using lexical-item and regularity data obtained from our data-preparation methods, our analysis revealed that the numerical values assigned to the lexical items by our revised homogeneity analysis (Gifi 1990; de Leeuw and Mair 2009; Mair and de Leeuw 2022; Ono and Fukazawa 2025) differed significantly. This difference was observed when

comparing the results of the analysis applied only to lexical-item data with those of the analysis used on data combining both lexical items and regularity.

The two classification results—one from lexical items alone and the other from both lexical items and regularities—offered significant insights into historical divergence and language contact in Ainu dialects, respectively, thereby validating our data-preparation methods.

Specifically, the classification result from data compiling both lexical items and regularity showed that (1) the Asahikawa, Nayoro, and Soya dialects formed an intermediate group relative to the southwestern and northeastern groups of Hokkaido dialects (cf. Ono 2019) and (2) they were positioned near the origin of the 3D coordinates calculated in homogeneity analysis. In homogeneity analysis, the origin of the coordinate system represents the position of a hypothetical dialect with the average linguistic characteristics of all dialects included in the analysis.

Thus, the Asahikawa, Nayoro, and Soya dialects are considered close to the “average” dialect among all those in Asai (1974). From a statistical perspective, this positioning indicates an area of historical language contact or mixture with various other dialects.

From the classification results from lexical items, it was found that (3) the Sakhalin dialects were divided into northern and southern groups and (4) the north Sakhalin group exhibited similarity to the Saru-Chitose dialect group. The latter finding aligns with what has been referred to in previous research as the “Saru-Chitose-Sakhalin-type” dialect (Nakagawa 1996).

Furthermore, the northern Sakhalin group is close to the Saru-Chitose dialect group. Conversely, the southern Sakhalin group is close to the northeastern Hokkaido dialect group, forming an A–B–A distribution from a geolinguistic viewpoint. This finding can hold significant implications for future historical research on the Ainu language.

Therefore, our proposed framework—which combines our data-preparation method with a revised homogeneity analysis applied to the resulting data—provides an alternative approach for maximizing the use of historical information in languages or dialects from lexicostatistical documents. In this paper, this approach can also capture and integrate the two main perspectives in linguistics: the geolinguistics view can correspond to the data of lexical items alone, the historical linguistics view can correspond to the regularity data, and geolinguistics and historical linguistics can potentially be integrated into the data compilation of both lexical items and regularity.

2. Background on lexicostatistical research in Ainu dialects

This section primarily aims to summarize previous lexicostatistical studies of Ainu dialects and to explain our motivation for developing new data-preparation methods in this study.

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).

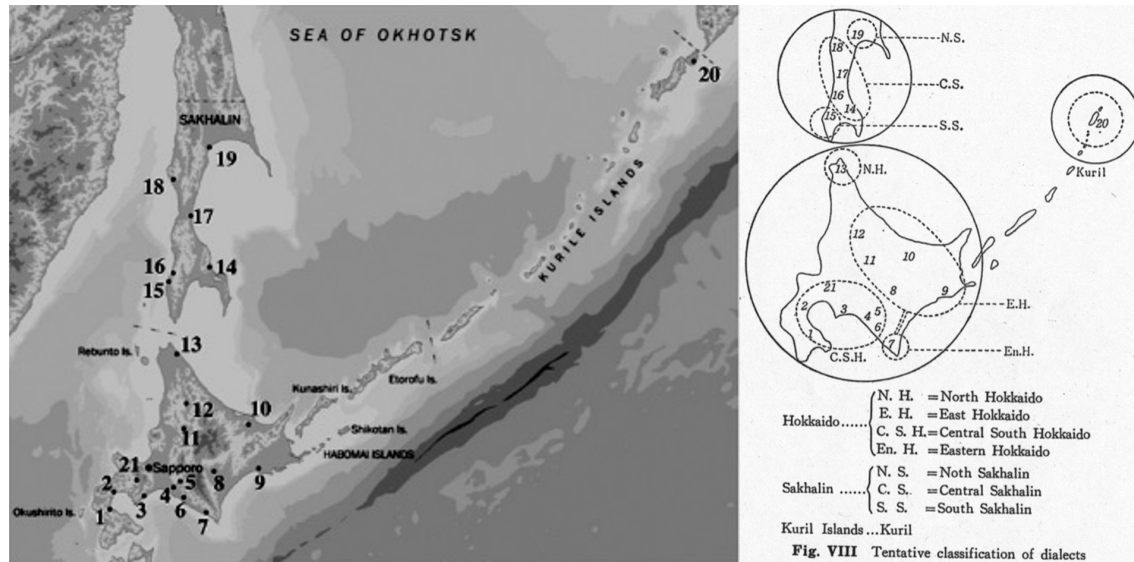


Figure 1. Map of Ainu dialects (Geospatial Information Authority of Japan, 2019). Left: Map of Ainu dialects (Geospatial Information Authority of Japan, 2019). Note: 1, Yakumo; 2, Oshamambe; 3, Horobetsu; 4, Biratori; 5, Nukibetsu; 6, Niikappu; 7, Samani; 8, Obihiro; 9, Kushiro; 10, Bihoro; 11, Asahikawa; 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). In this study, southern Sakhalin is called Sakhalin according to the conventional usage in Hattori and Chiri (1960) and Asai (1974).

Their study investigated nine-teen Ainu dialects from that time, as many dialects were on the verge of disappearing. The second study, Asai (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).

Both Hattori and Chiri (1960) and Asai (1974) classified Ainu dialects using statistical analyses of basic vocabulary (200 and 202 items, respectively). However, these classifications could not be directly compared until recent studies (Ono 2020; Ono and Fukazawa 2023, 2024). The primary hindrance was that, while Hattori and Chiri (1960) documented their methodology, Asai (1974) did not clearly indicate his selection process

for lexical items or the specific cognacy judgments adopted for those items.¹

Using the computational algorithms in Ono (2020), the limited descriptions of cognacy judgments in Asai (1974), and the basic hierarchy of cognacy in Ainu dialects, previous studies (Ono and Fukazawa 2023, 2024) successfully identified 110 lexical items and their corresponding cognacy judgments in Asai (1974).

These studies demonstrated that Hattori and Chiri's (1960) and Asai's (1974) classifications were based on different lexical items and distinct cognacy judgments regarding lexical forms. This information retrieval has enabled researchers to compare and examine both classifications from the perspective of contemporary Ainu dialectology. Consequently, Asai's (1974) judgments were found to be finer than those of Hattori and Chiri (1960). For example, *cep* and *ceh* in the basic vocabulary of 'fish' are judged as cognates in Hattori and Chiri (1960), but are later considered as noncognates in Asai (1974).

However, previous studies still reported data-preparation issues (Fukazawa and Ono 2025; Ono and Fukazawa 2025). Ono and Fukazawa (2024) compared the two classifications using the same 110-word list—one adopted by Hattori and Chiri (1960) and the other adopted by Asai (1974)—and clarified the nature of classification in humanities data. They demonstrated that the inclusion relationship between Asai's (1974) judgments—which are finer than Hattori and Chiri's—and those of Hattori and Chiri (1960) did not lead to one classification that subsumed the other (i.e., Hattori and Chiri's classification being included in Asai's, or vice versa).

Instead, the two classifications in Figure 2, calculated by our statistical analyses, resulted in two different patterns of Ainu dialects: an eastern-western pattern from Asai's judgments and a Saru-Chitose type from Hattori and Chiri's judgments. As concluded by Ono and Fukazawa (2024), the classification in humanities data should be evaluated not by the cognacy judgments made in each study alone but by the consistency between the classification's purpose and those judgments.

Nonetheless, Ono and Fukazawa (2024) have simultaneously opened another fundamental question in the classification of Ainu dialects. Indeed, while classification results can vary and should be examined for consistency with the study's purpose and the judgments used, a more fundamental, unresolved issue in data preparation remains, that is, whether researchers can maximize the information potentially obtainable from

¹ There are several lexical forms that can be considered as sharing the same origin from the perspectives of Ainu linguistics in Table 1 of the main text: *hunna* and *hunat* in '6. who'; *ruyanpe* and *ruwanpe* in '76. rain' (the rain); *apto*, *ahto*, and *atto* in '76. rain' (the rain); *nekona* and *nekon* in '104. how'; *mak*, *manak*, and *maknak* in '104. how'; *hekaci* and *hekattar* in '151. child'; *ekaci* and *ekattar* in '151. child'; *rarak*, *raarak*, and *raarah* in '167. smooth'. In linguistics, these views are called as "cognacy judgements". For example, *hunna* and *hunat* are considered as "cognate." Several different cognacy judgements can hold on the same lexicostatistical documents that lead to different classifications of languages or dialects (Ono and Fukazawa 2023, 2024).

lexicostatistical documents—beyond current data-preparation methods—and, if possible, how to extract and computationally quantify this information as symbols.

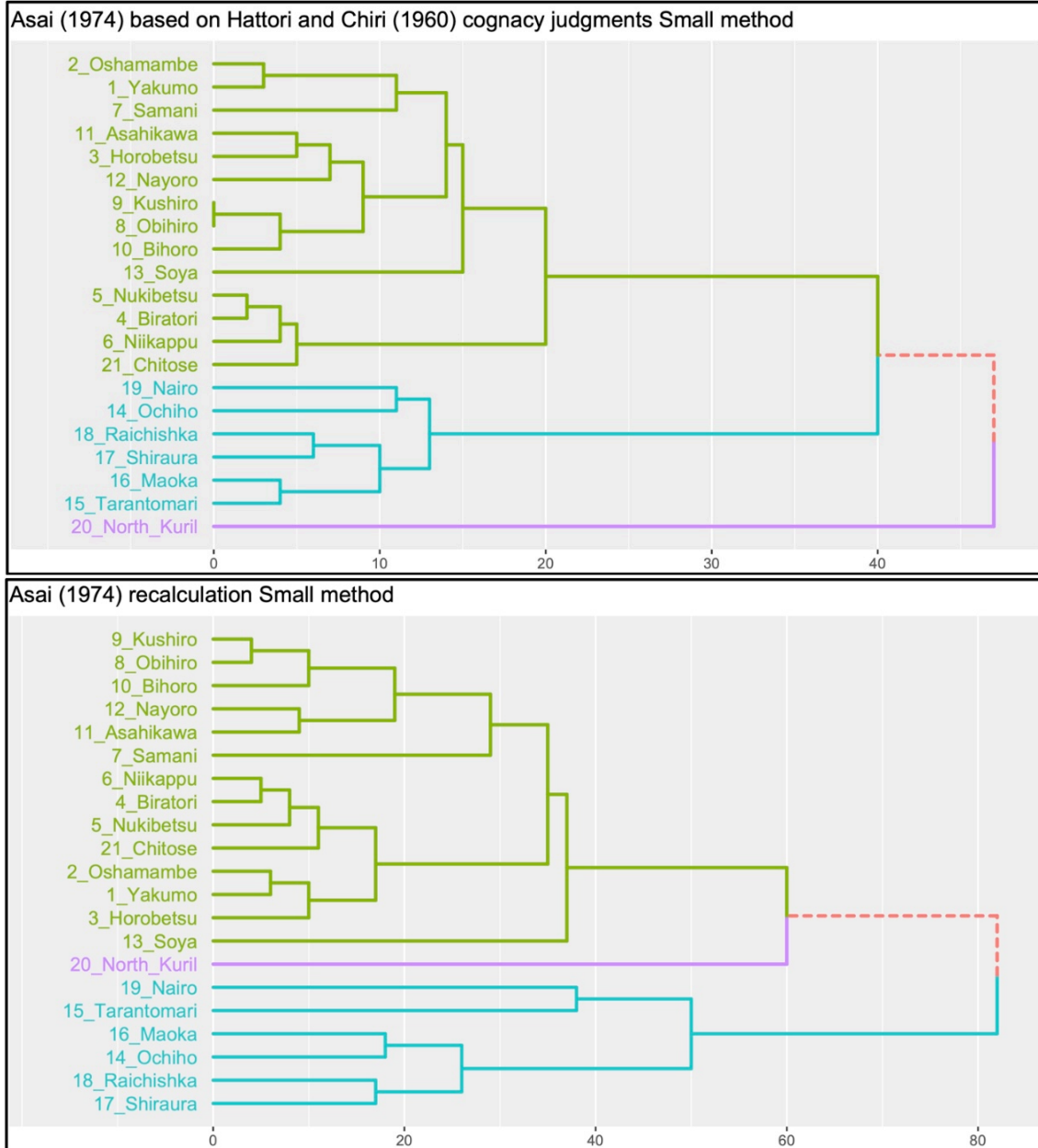


Figure 2. The two comparable classifications of Ainu dialects in Ono and Fukazawa (2024: 108–109). Top: The classification of Ainu dialects obtained using cluster analysis based on Hattori and Chiri (1960)’s cognacy judgments; Bottom: The classification of Ainu dialects obtained using cluster analysis based on Asai’s (1974) cognacy judgments.

The scope of lexicostatistics originally focused on historical information about languages or dialects. Still, the methodological limitation of data preparation can restrict

its scope to “classification,” as some linguistic patterns embedded across multiple lexical items may be amplified in classification results during statistical quantification as discussed in Section 2.2 of the main text.

Therefore, whether the fullest information in Ainu dialectology can be extracted from Asai’s documents and whether the historical divergence or language contact is visualized in Ainu dialects, obscured by data-preparation methods in previous lexicostatistics, remain unclear.

3. Quantification Methods and Their Linguistic Meanings

The Section 3 focuses on our revised homogeneity analysis, a numerical quantification method that assigns numerical values to the symbols extracted during our data preparation. A key advantage of homogeneity analysis will also be discussed: the mathematical property whereby the sum of the values assigned to lexical forms within each lexical item and rules within each regularity is approximately equal to the quantified values of each dialect.

Table 1. 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 (*konru*, RU(P/H), *tup*) and rules ($-p/t/k$ 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

This method enables researchers to verify the extent to which specific lexical forms or rules influence the dialect configuration. Furthermore, it identifies which regularities contribute to changes in relative differences observed between the two dialect configurations—specifically, between the configurations obtained from the lexical dataset alone and the one derived from the combined dataset of lexical items and regularities. Additionally, quantification procedures of homogeneity analysis will be interpreted within a linguistic context.

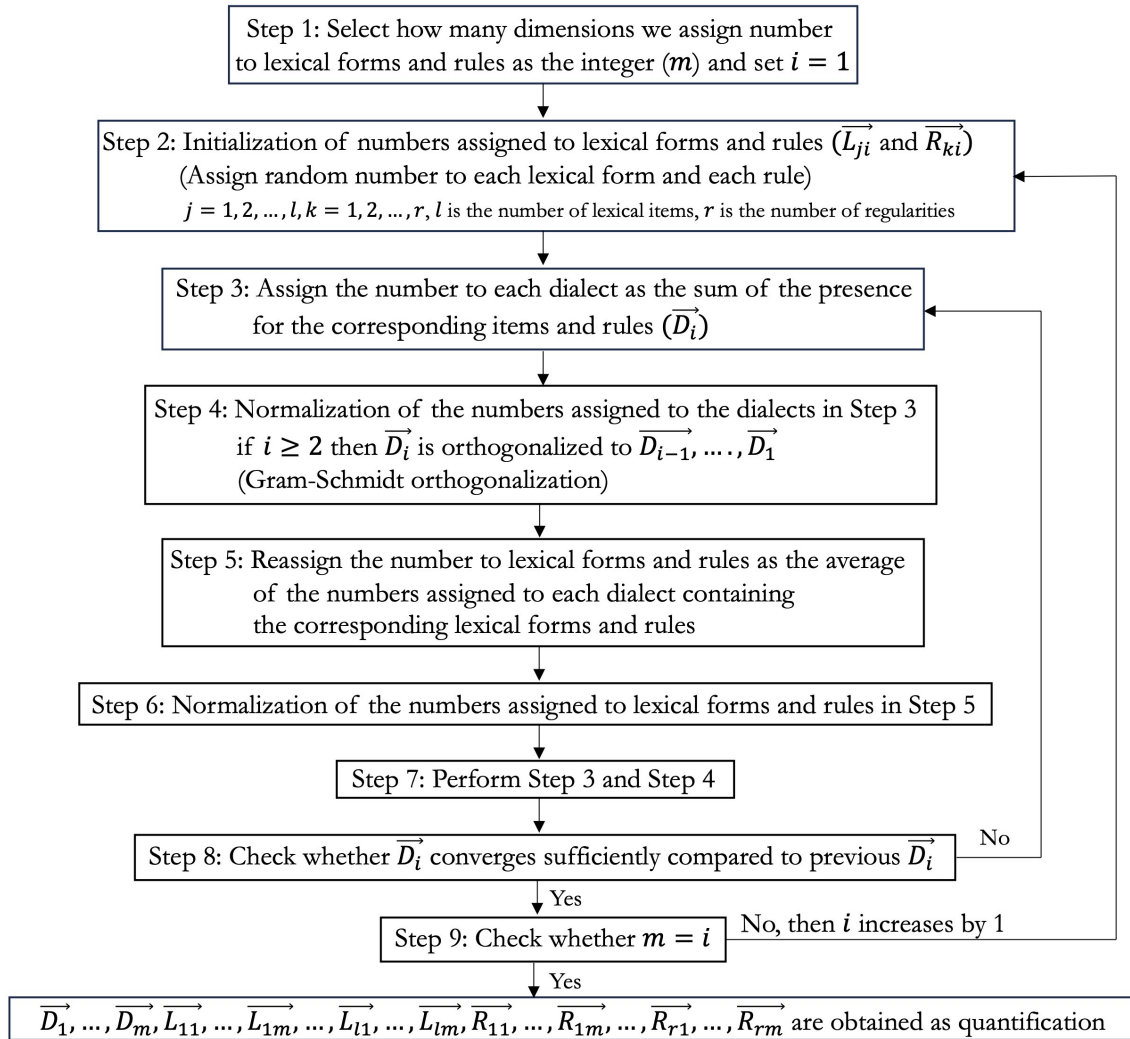


Figure 3. Flowchart of homogeneity analysis applied to the dataset, combining both lexical items and regularity from the perspective of the method of reciprocal averages. Note that the flowchart of homogeneity analysis applied to only lexical-item data corresponds to the flowchart without the information of the vector of R .

3.1 Specific Procedures in Homogeneity Analysis and the Quantified Configuration of Dialects

Table 1 is the example of lexical-item data and regularity data extracted by our data-preparation methods. The computational algorithm of homogeneity analysis (Ono and Fukazawa 2025), can be explained via the method of reciprocal averages (MRA; Richardson and Kuder 1933; Horst 1935). The simplified algorithm for homogeneity analysis, which focuses only on quantifying dialects, lexical forms, and rules, rather than lexical items and regularities, is summarized in Figure 3.

Let us focus on the specific steps in Figure 3, using the example data from Table 1. In Step 1, the number of dimensions to which we assign numbers to lexical forms and rules is set to 3 (i.e., $m = 3$ in Figure 3) for the sake of 3D visualization of Ainu dialects. The value of i is set as 1 in this step.

In Step 2, a random number is assigned to the vector of L11 (consisting of L1_1, L1_2, L1_3). The vector of R11 (consisting of R1_1, R1_2), where L1_1, L1_2, L1_3, R1_1, R1_2 correspond to the assigned number to *konru*, RU(P/H), *tup*, $-p/t/k$ and $-h/s$ respectively, and L1_1 = 0.8, L1_2 = 1.4, L1_3 = 1.8, R1_1 = 0.6, R1_2 = 1.0, are assumed to be assigned tentatively.

In Step 3, we assign a number to each dialect as the sum of the presence for the corresponding lexical items and rules. For example, the Yakumo dialect contains *konru* in lexical item 1 and $-p/t/k$ in regularity 1. Thus, D1 is assigned as $L1_1 + R1_1 = 0.8 + 0.6 = 1.4$. The same calculation applies to D2–D21. Thus, we define the vector of D1 consisting of D1, D2, ..., D21. In our case, D1–D11 = 1.4, D12 = 2.8, D13 = 2.0, D14 = 2.4, D15 = 2.0, D16–D18 = 2.4, D19 = 2.4, D20–D21 = 1.4 are assigned to each dialect.

In Step 4, the vector of D1 is normalized by (1) first subtracting the average value of D1, D2, ..., D21 from each value of D1, D2, ..., D21, where the sum of each transformed value of D1', D2', ..., D21' is set as 0, and then by (2) dividing each transformed value of D1', D2', ..., D21' by the length of the transformed vector of D1, that is, $\sqrt{D1'^2 + D2'^2 + \dots + D21'^2}$, where the transformed vector of D1, D1', shows the length as 1. In our case, D1–D11 = -0.16, D12 = 0.47, D13 = 0.11, D14 = 0.29, D15 = 0.11, D16–D18 = 0.29, D19 = 0.29, D20–D21 = -0.16.²

²If the value of i is 2 and the assign value of D1, D2, ..., D21 has been calculated and determined as specific values (i.e., the complete vector of D1), the additional normalization is applied to the vector of D2, as the transformed vector of D2, D2' equal to the vector of D2 - (inner product between the vector of D2 and the complete vector of D1) over (inner product between the complete vector of D1 and the complete vector of D1) times the complete vector of D1. The procedure is called Gram-Schmidt orthogonalization, which projects the vector D2 onto directions that are orthogonal to D1. In other words, the completed vector of D2 does not have any information related to the complete vector of D1 or orthogonalized to the complete vector of D1 from the mathematical viewpoint. The Gram-Schmidt orthogonalization will also work when $i = 3$, provided the values of D1, D2, ..., D21 have been calculated and assigned specific values in the first two dimensions (i.e., the complete vector of D1 and the complete vector of D2). The additional normalization is applied to the vector of D3, as the transformed vector of D3, D3' equal to the vector of D3 - (inner product between the vector of D3 and the complete vector of D1) over (inner product between the complete vector of D1 and the complete vector of D1) times the complete vector of D1 - (inner product between the vector of D3 and the

In Step 5, we reassign the number to lexical forms and rules as the average of the numbers assigned to each dialect containing the corresponding lexical forms and rules. For example, D1–D12, D20, and D21 contain *konru*, and L1_1 is the average of D1–D12, D20, and D21, equal to -0.12 . The same calculation applies to L1_2, L1_3, R1_1, and R1_2, respectively. In our case, L1_1 = -0.12 , L1_2 = 0.26 , L1_3 = 0.29 , R1_1 = -0.07 , R1_2 = 0.29 .

In Step 6, the vectors L11 and R11 are normalized so that each vector is approximately zero and has length 1, using calculations similar to those in Step 4. In our case, L1_1 = -0.82 , L1_2 = 0.37 , L1_3 = 0.45 , R1_1 = -0.71 , R1_2 = 0.71 .

In Step 7, Steps 3 and 4 were performed again. In our case, D1–D11 = -0.15 , D12 = -0.07 , D13 = 0.11 , D14 = 0.41 , D15 = 0.11 , D16–D18 = 0.41 , D19 = 0.12 , D20–D21 = -0.15 .

In Step 8, the two vectors of D1 are compared,³ and the sum of squared differences is calculated and checked whether the sum of squared differences is sufficiently small, because the mathematical and statistical theory of homogeneity analysis guarantees the vector of D1 in the Step as converged to the complete vector of D1 with the criterion using a sufficiently small value of the sum of squared differences. In our case, the difference between the two vectors of D1 and the nonnegligible sum of squared differences can be easily understood. Thus, we return to Step 3. In the other case, if we can consider the two vectors of D1 show sufficiently small value of the sum of squared differences, then we save the vector of D1 in Step 7 as the complete vector of D1 (i.e., the quantified scores of dialects), the corresponding vector of L11 and R11 as the full vector of L11 and R11.

We then set i to 2 and perform Step 2 to calculate the complete vector of D2, L12, and R12. After calculating the complete vectors D2, L12, and R12, we set i to 3 and perform to Step 2 to calculate the complete vectors D3, L13, and R13. After calculating the complete vectors for D3, L13, and R13, the homogeneity analysis will return the complete vectors for D1, D2, D3, and L11, and L12, L13, R11, R12, and R13 for the quantification results.⁴

Using the complete vector of D1, D2, and D3, the configuration of Ainu dialects is

complete vector of D2) over (inner product between the complete vector of D2 and the complete vector of D2) times the complete vector of D2. The Gram–Schmidt orthogonalization projects the vector of D3 into a direction that is not related to the direction of the complete vector of D1 or that of the complete vector of D2. In other words, the completed vector of D3 does not have any information related to the complete vector of D1 or the complete vector of D2, or orthogonalized to the complete vector of D1 and the complete vector of D2, from the mathematical viewpoint.

³ -0.16 , D12 = 0.47 , D13 = 0.11 , D14 = 0.29 , D15 = 0.11 , D16–D18 = 0.29 , D19 = 0.29 , D20–D21 = -0.16 and D1–D11 = -0.15 , D12 = -0.07 , D13 = 0.11 , D14 = 0.41 , D15 = 0.11 , D16–D18 = 0.41 , D19 = 0.12 , D20–D21 = -0.15 .

⁴In another quantification procedure, where we focus on the vector of lexical items (i.e., the vector of L1), the other part of the homogeneity analysis works in the same procedure as in Figure 5.

illustrated from the lexical-item data and the compiled dataset of both lexical-item data and regularity data in 3D space. We can also understand that the disposition of dialect in the first, second, and third dimensions corresponds to the complete vector of D1, D2, and D3 in the lexical-item data.⁵

3.2 Linguistic Meaning of Quantification Procedures in Homogeneity Analysis

The previous part explained the specific calculation procedures for homogeneity analysis using the example data shown in Figure 2. In this subsection, the linguistic meaning of quantification procedures in homogeneity analysis and their innovative features in linguistic research are explained, as outlined in the following six points.

First, as illustrated in Step 3 of Figure 5, the quantification scores for dialects (D1–D21) essentially represent the aggregation (or sums) of the presence of corresponding lexical forms and rules. This characteristic empowers researchers to identify precisely which lexical forms or rules determine the dialectal configuration along specific dimensions (e.g., the first, second, or third dimension). Furthermore, it allows us to analyze how the inclusion of regularity data alters these configurations by comparing the results as discussed in Section 5. This direct interpretive procedure is not readily achievable by other quantification methods, such as correspondence analysis (Benzécri et al., 1973), which is commonly used in geolinguistics. In this regard, our recent study (Ono and Fukazawa 2025) first demonstrates the potential of homogeneity analysis in the linguistic domain.

Second, the normalization process in Step 4 of Figure 5 transforms the dialect values (D1–D21) so that their mean is zero. This guarantees that the zero value corresponds to the position of a “virtual dialect” possessing the average characteristics of all dialects in the corresponding dimension. While actual dialects do not necessarily occupy this exact zero position, the origin point in the 3D configuration (as illustrated in Section 5) represents the location of this virtual dialect across all dimensions (i.e., all three dimensions in our case). In other words, dialects positioned near the origin in the 3D configuration—whether derived from lexical-item data or the combined dataset—can be considered as those exhibiting the average characteristics of all dialects in the data.

Third, the Gram–Schmidt orthogonalization (Gram 1883; Schmidt 1907) in Step 4 eliminates vector directions of the dialects calculated in previous dimensions, ensuring that the new direction is orthogonal to those previously estimated. This enables researchers to visualize historical information in lexicostatistical documents that is obscured by primary or secondary historical details and to easily capture it through human recognition, given its high variation. Thus, homogeneity analysis can potentially reveal

⁵Note that the quantification results of D1, D2, D3, L11, L12, and L13 are different between the lexical-item data and the compiling dataset of both lexical-item data and regularity data.

undiscovered information in lexicostatistical documents, such as historical divergence or language contact in our case.

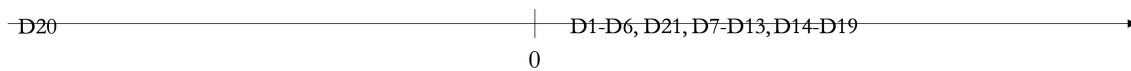
Fourth, the final dialect vector (e.g., D1 in the first dimension) is calculated by maximizing the variation of quantified scores (D1–D21), using the criterion of the sum of squared differences from the mean. However, this criterion has a limitation: it cannot inherently distinguish between different distributional patterns, as illustrated in Figure 6.

Case 1 (top of Figure 4) shows a scenario in which only one dialect (D20) is strongly quantified in the negative direction, whereas the others are clustered in a small positive range.

Case 2 (bottom of Figure 4) shows a different distribution: the southwestern Hokkaido (D1–D6) and Sakhalin (D14–D19) groups are negative, the intermediate group (D11–D13) is positioned around the center (near zero), the northeastern Hokkaido group (D7–D10) is positive, and the Northern Kuril dialect is strongly positive.

We must note that the calculated vector does not necessarily capture a linguistically meaningful dimension. Since homogeneity analysis assigns numbers to maximize the sum of squared differences (variance), the resulting mathematical maximization does not always correspond to a linguistically interpretable direction. For example, Ono and Fukazawa (2025) have identified linguistically significant directions by combining three dimensions derived from a homogeneity analysis of information extracted from Asai's lexicostatistical documents, based on Asai's cognacy judgments (Fukazawa and Ono 2025).

Case 1



Case 2

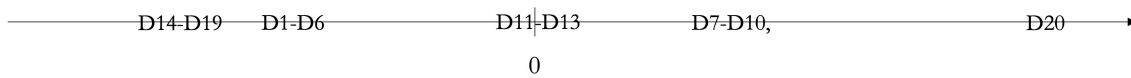


Figure 4. Two quantification results show the same error defined as squared differences around mean. Top: Northern Kuril (D20) is quantified negatively as an outlier, and the other dialects are quantified positively within small range; Bottom: southwestern Hokkaido group (D1–D6) and Sakhalin group (D14–D19) are quantified negatively, the intermediate group (D11–D13) is positioned around the center point (i.e., around the value of 0), northeastern Hokkaido group (D7–D10) is quantified positively, and the Northern Kuril (D20) is quantified strongly in positive direction.

Fifth, the quantification process in homogeneity analysis is formalized as an eigenvalue problem. This ensures that the first dimension captures the direction of maximum variation. Subsequent dimensions are derived from being orthogonal to the preceding ones, capturing the remaining variation in descending order (i.e., the second-

largest, third-largest, etc.). Consequently, this method allows researchers to investigate subtler structures (finer structures) in lexicostatistical documents by examining the lower dimensions. Since these dimensions are orthogonalized via the Gram–Schmidt process and exhibit minor variations, as dictated by eigenvalue theory, they reveal information distinct from the dominant trends in higher dimensions.

In the next section, we illustrate the quantified scores for lexical forms and rules obtained through homogeneity analysis, and we examine the visualization results from an Ainu linguistic perspective. Here, the explanation of specific procedures in this subsection will help distinguish the findings in visualizations that result from the calculation process itself from those that cannot be considered byproducts of the calculation process but rather purely linguistic findings.

4. The Quantified Scores for Lexical Forms and Rules

This section illustrates the quantified scores for lexical forms and rules in all 21 Ainu dialects in Figure 1, by applying our revised homogeneity analysis to the two datasets: one is lexical item data, and the other combines both lexical item and regularity data. On Figures 5–67, the quantified scores derived from lexical-item data (blue graphs) are markedly altered by the addition of regularity data (red graphs), respectively.

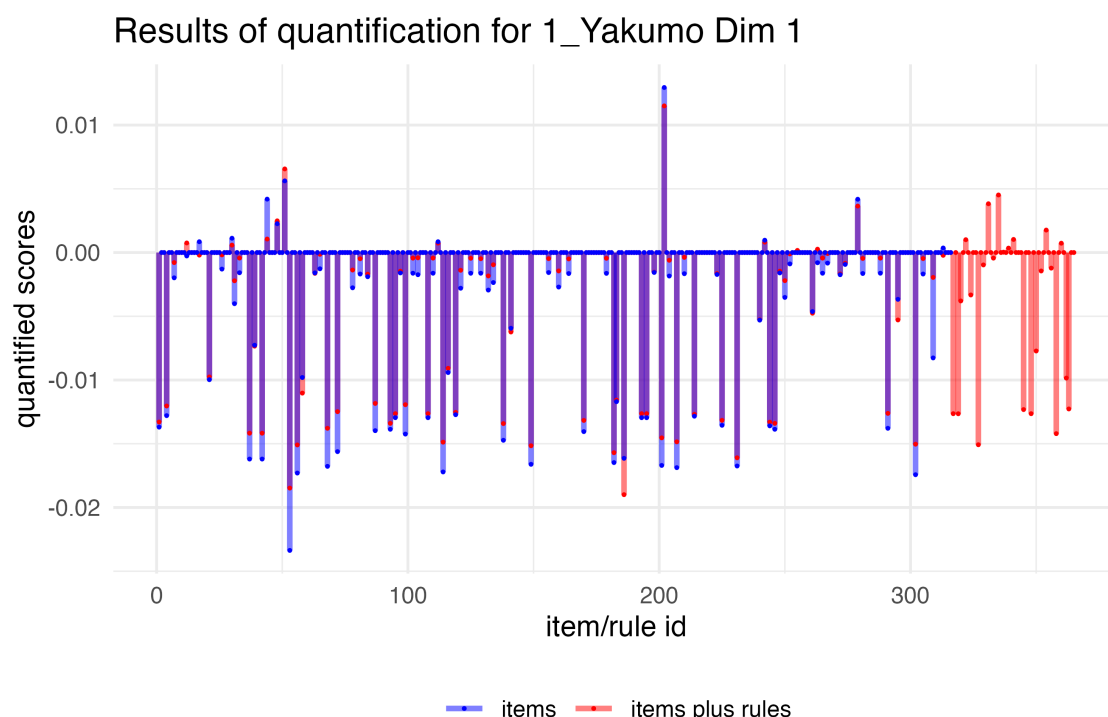


Figure 5. The quantified scores in the first dimension of the Yakumo dialect (No. 1 in Figure 1).

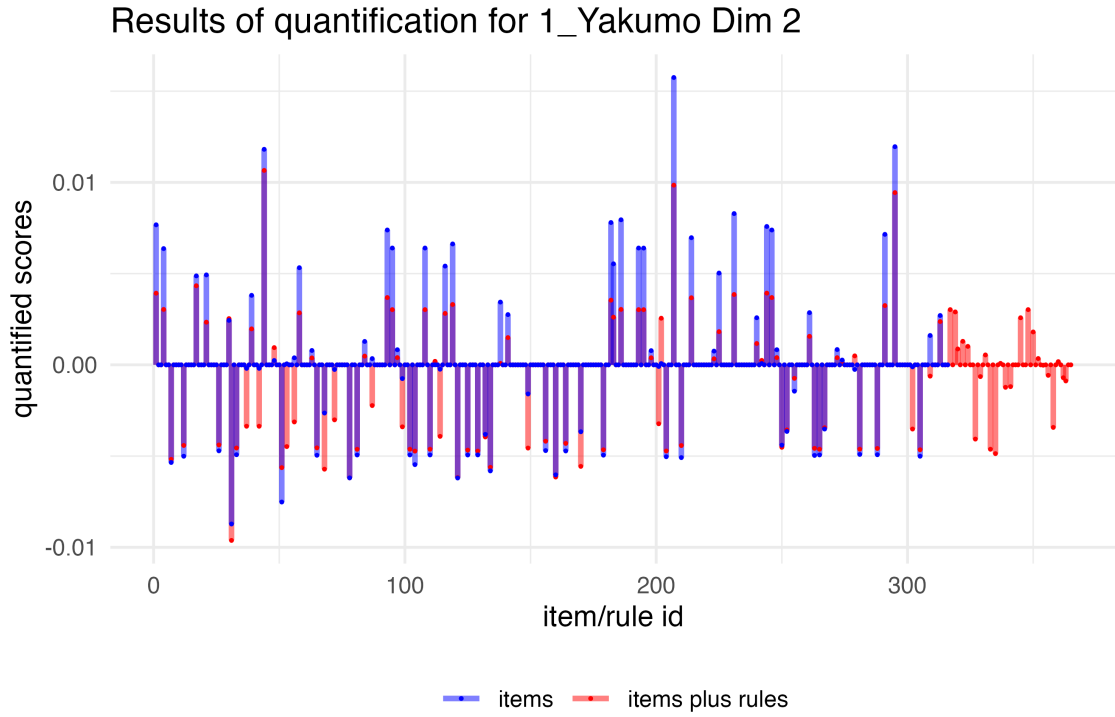


Figure 6. The quantified scores in the second dimension of the Yakumo dialect (No. 1).

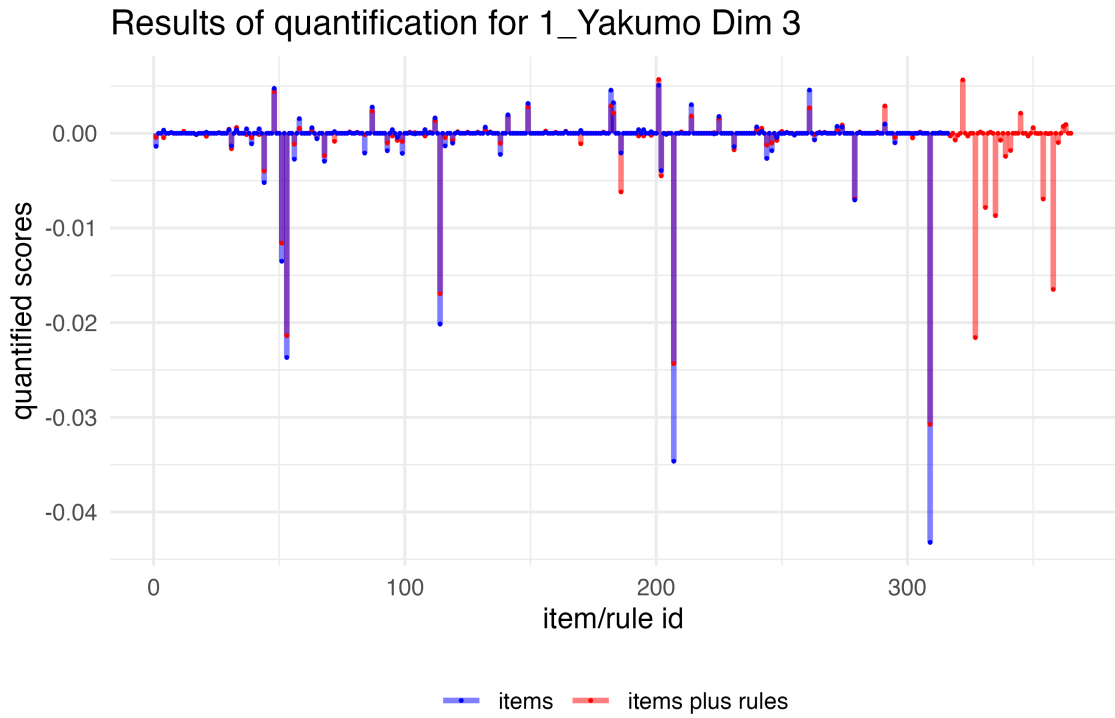


Figure 7. The quantified scores in the third dimension of the Yakumo dialect (No. 1).

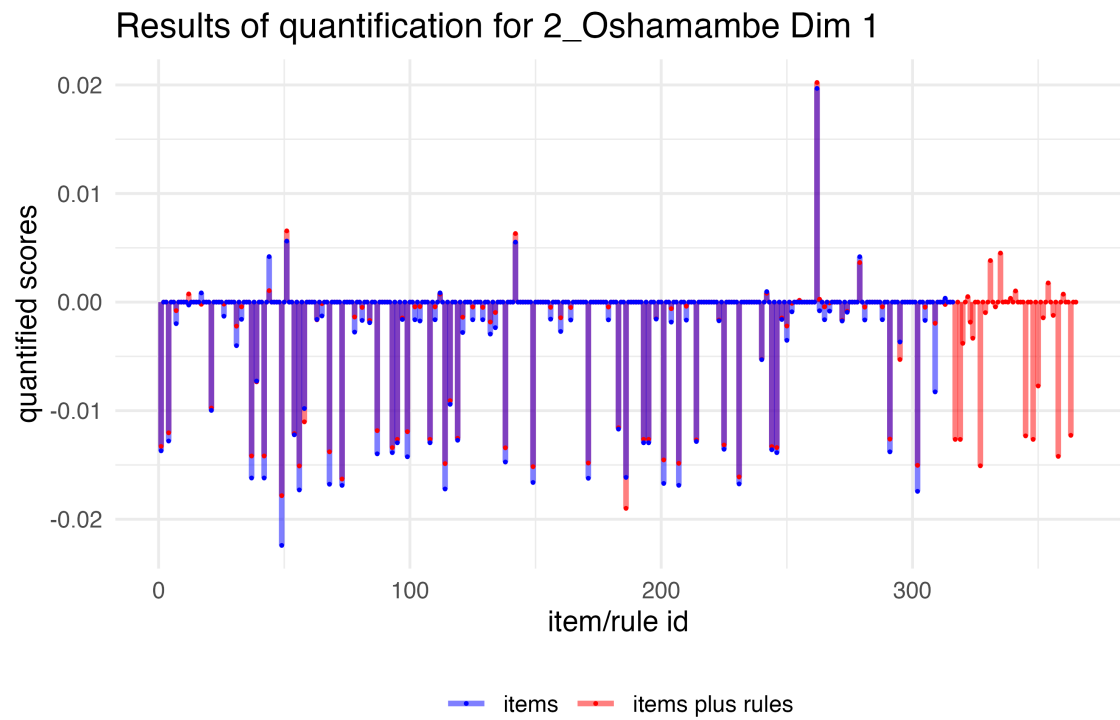


Figure 8. The quantified scores in the first dimension of the Oshamambe dialect (No. 2 in Figure 1).

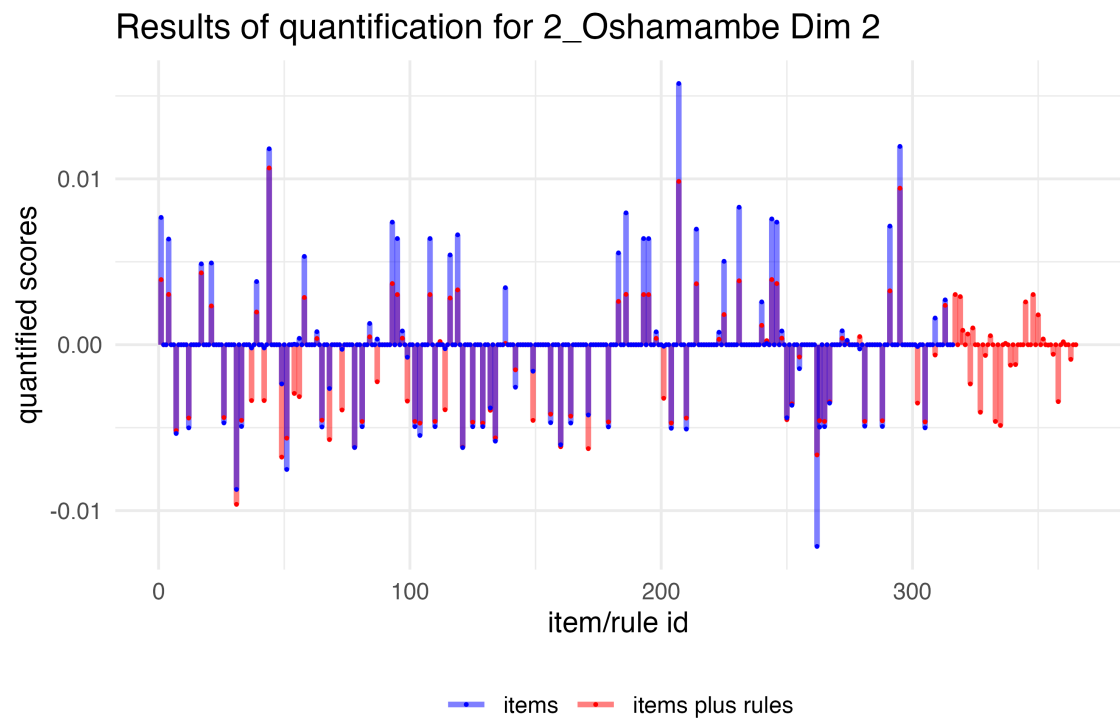


Figure 9. The quantified scores in the second dimension of the Oshamambe dialect (No. 2).

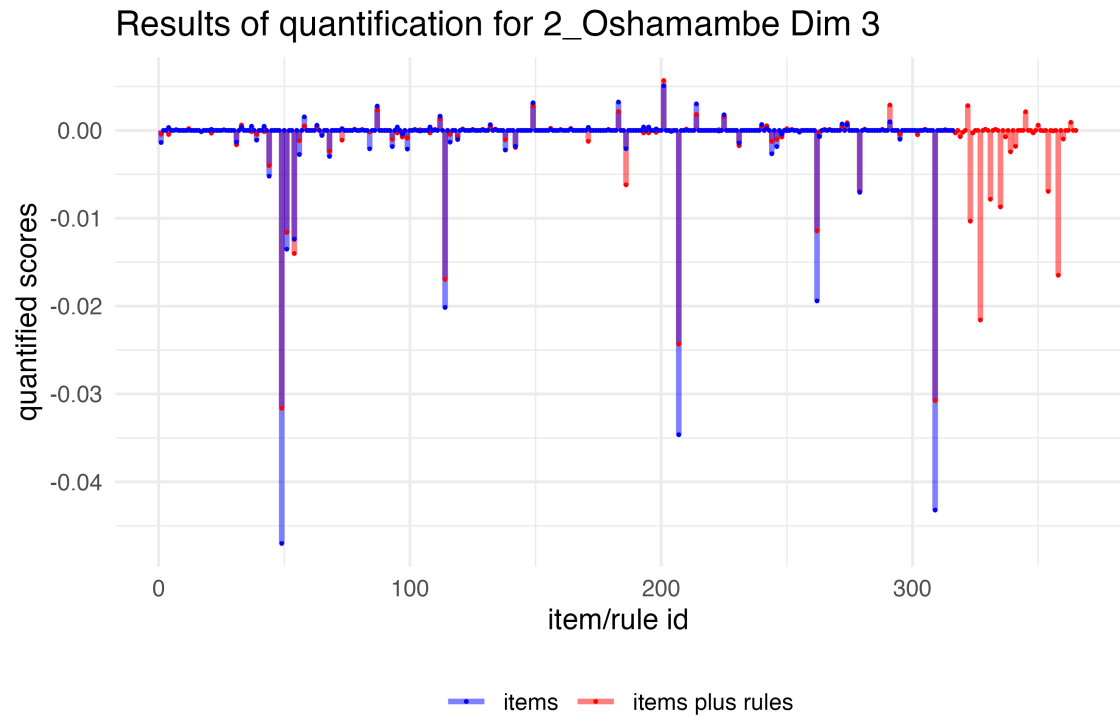


Figure 10. The quantified scores in the third dimension of the Oshamambe dialect (No. 2).

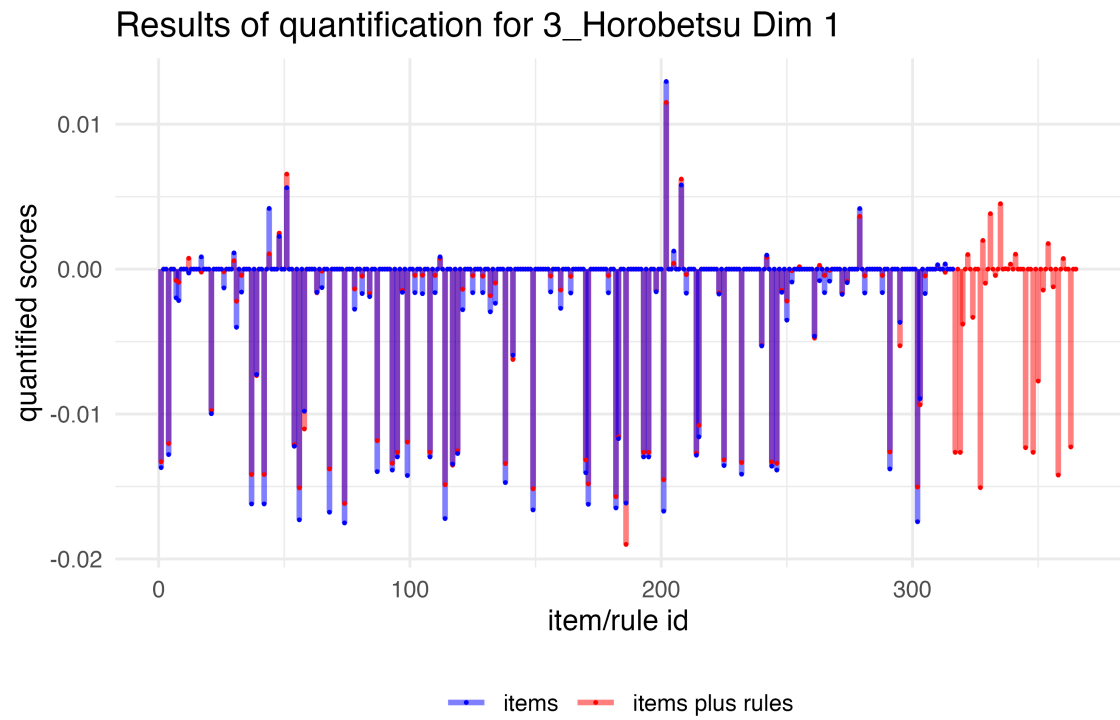


Figure 11. The quantified scores in the first dimension of the Horobetsu dialect (No. 3 in Figure 1).

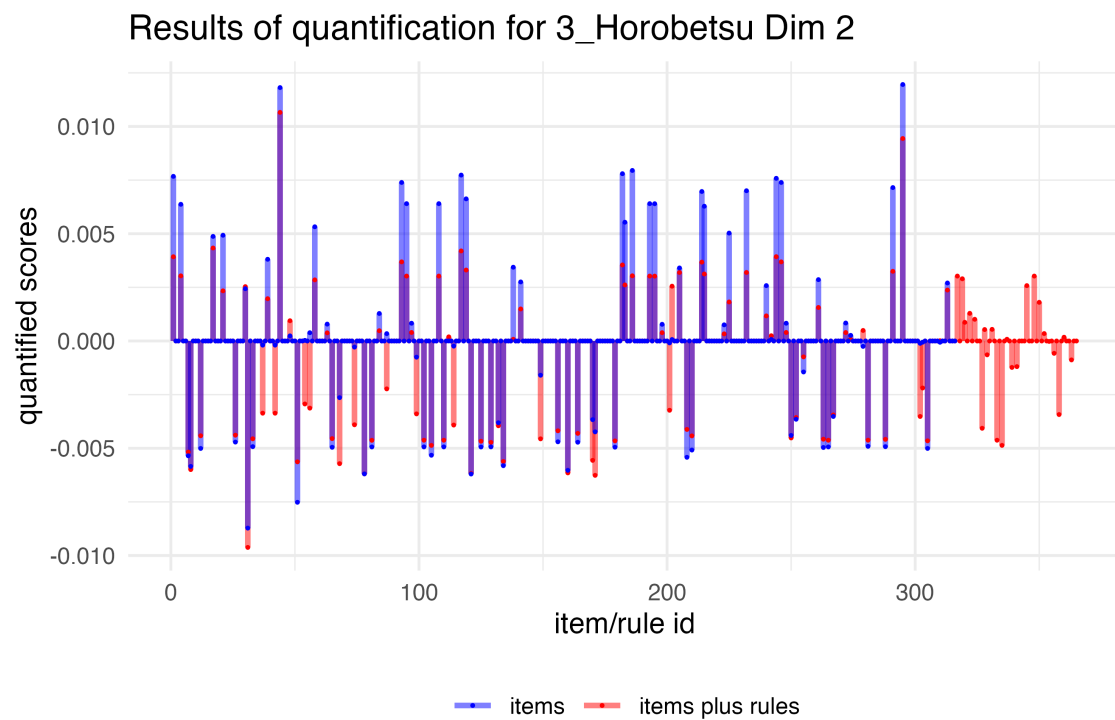


Figure 12. The quantified scores in the second dimension of the Horobetsu dialect (No. 3).

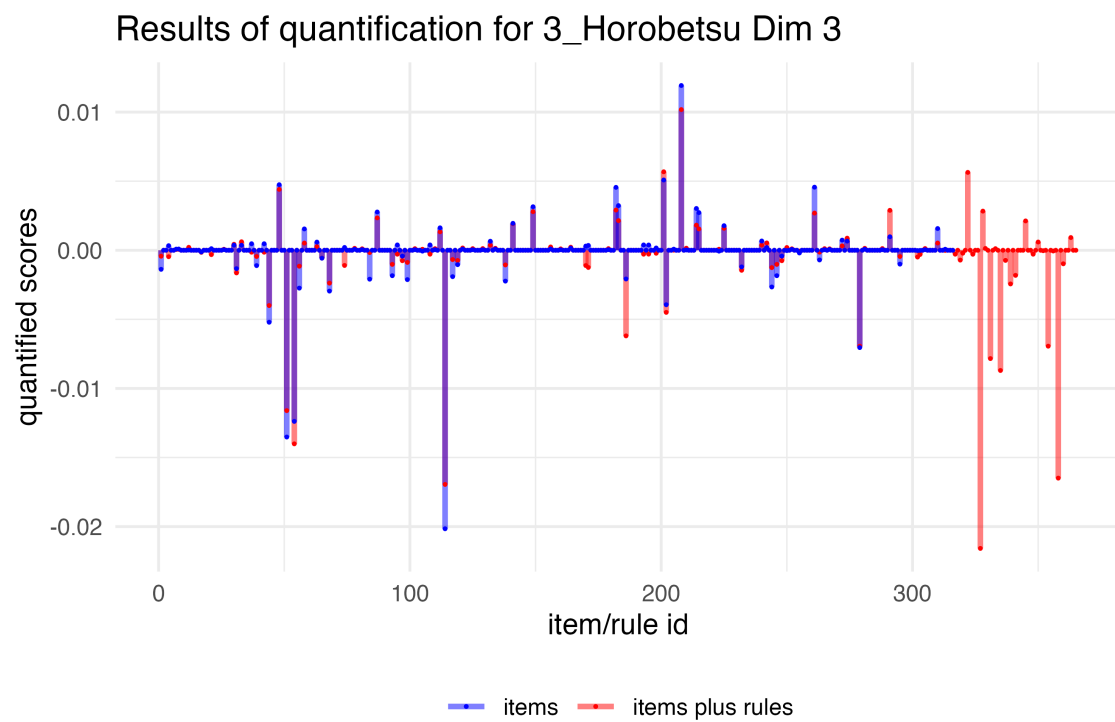


Figure 13. The quantified scores in the third dimension of the Horobetsu dialect (No. 3).

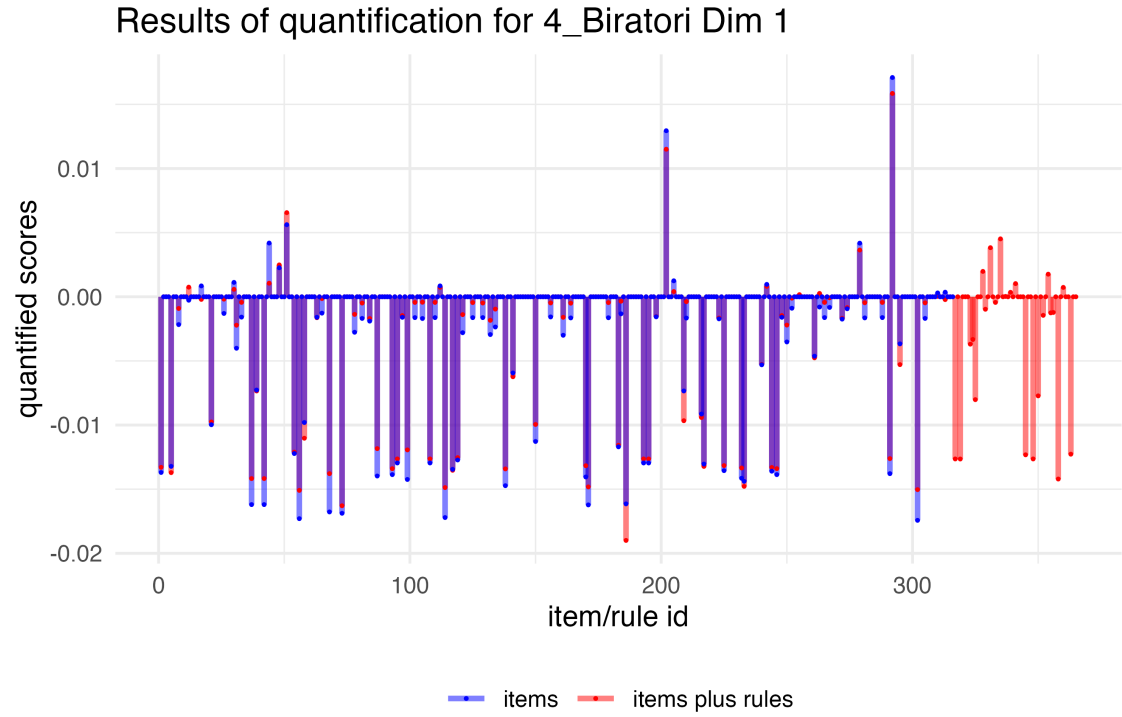


Figure 14. The quantified scores in the first dimension of the Biratori dialect (No. 4 in Figure 1).

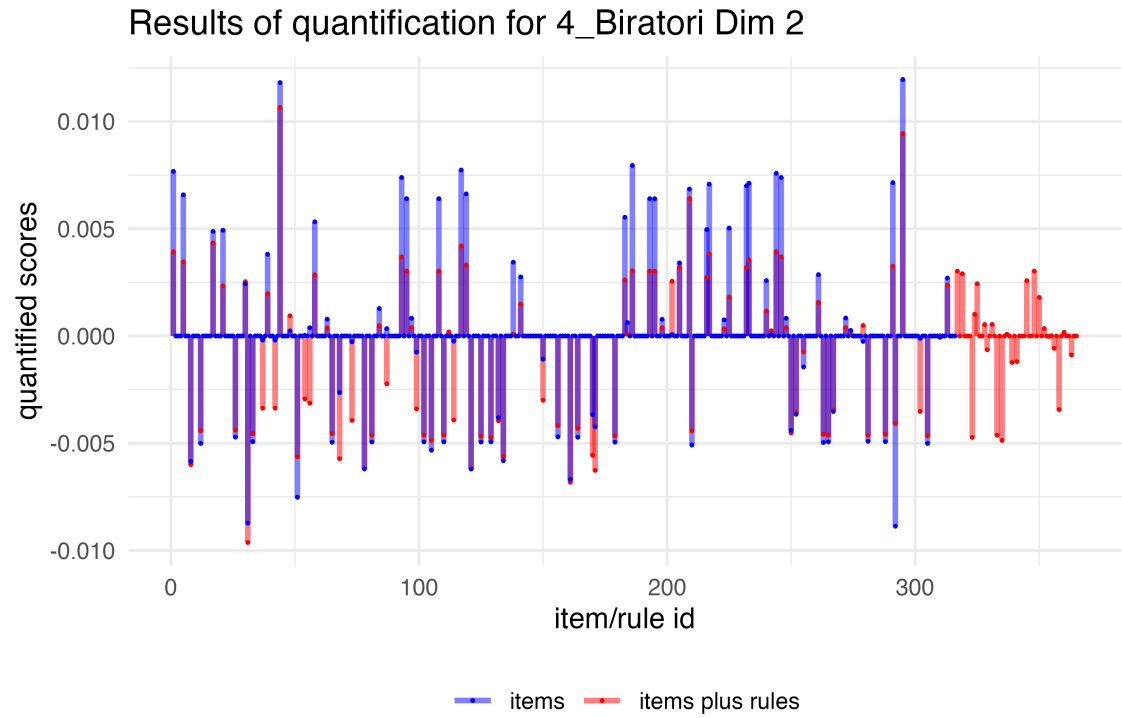


Figure 15. The quantified scores in the second dimension of the Biratori dialect (No. 4).

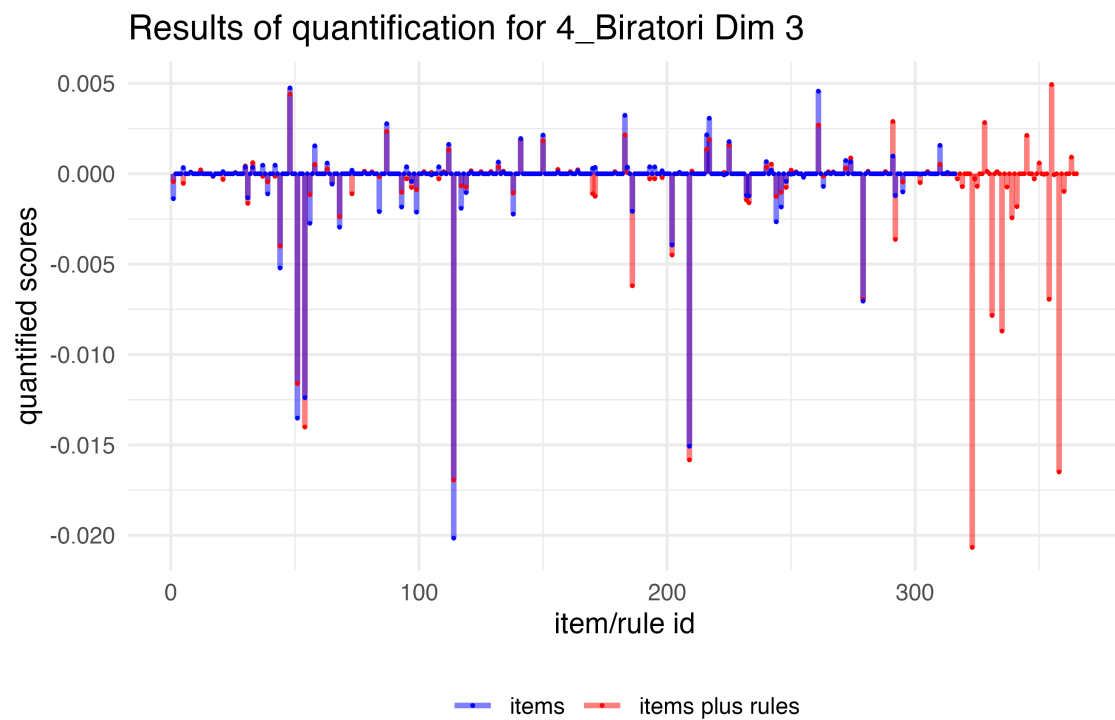


Figure 16. The quantified scores in the third dimension of the Biratori dialect (No. 4).

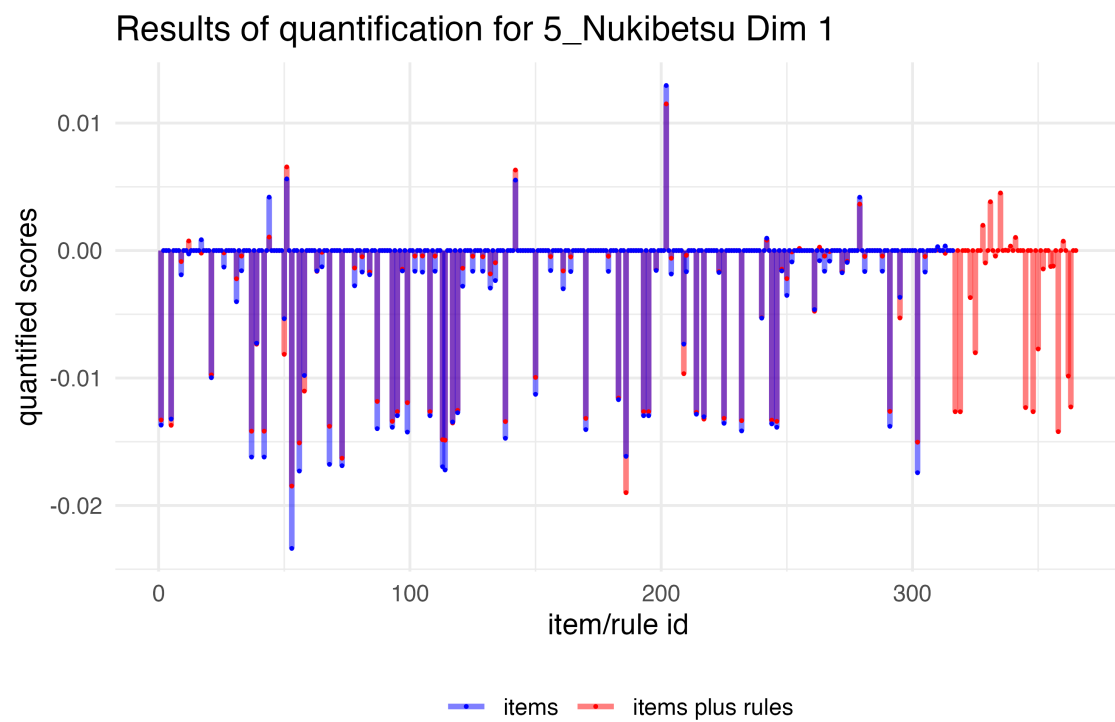


Figure 17. The quantified scores in the first dimension of the Nukibetsu dialect (No. 5 in Figure 1).

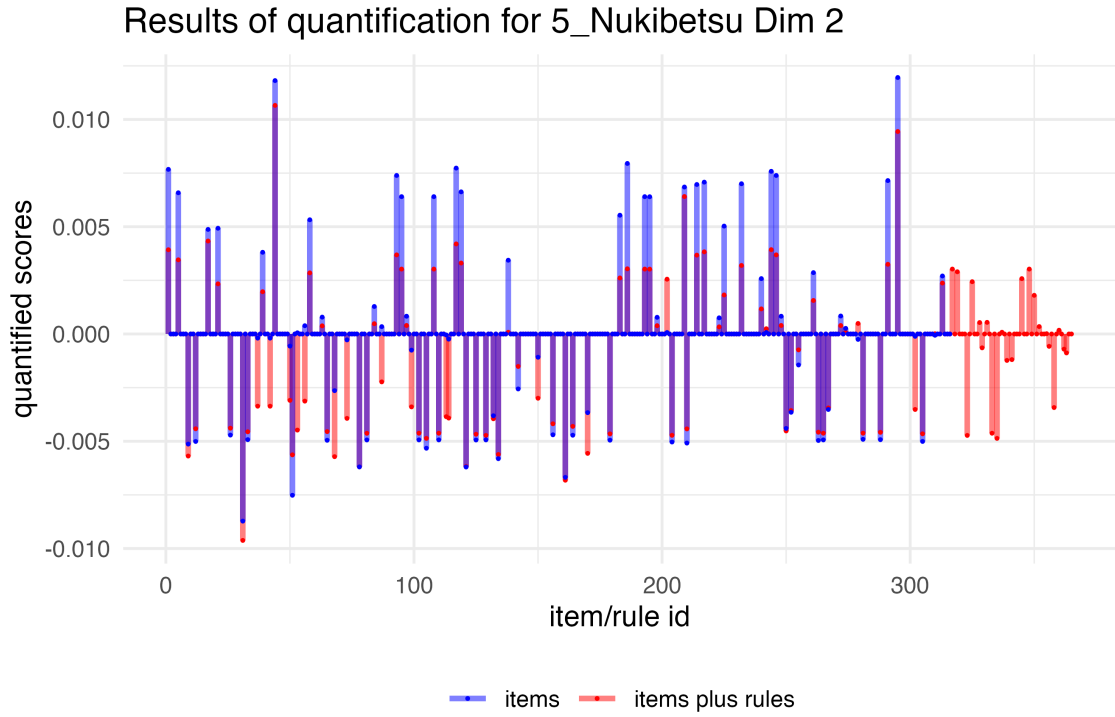


Figure 18. The quantified scores in the second dimension of the Nukibetsu dialect (No. 5).

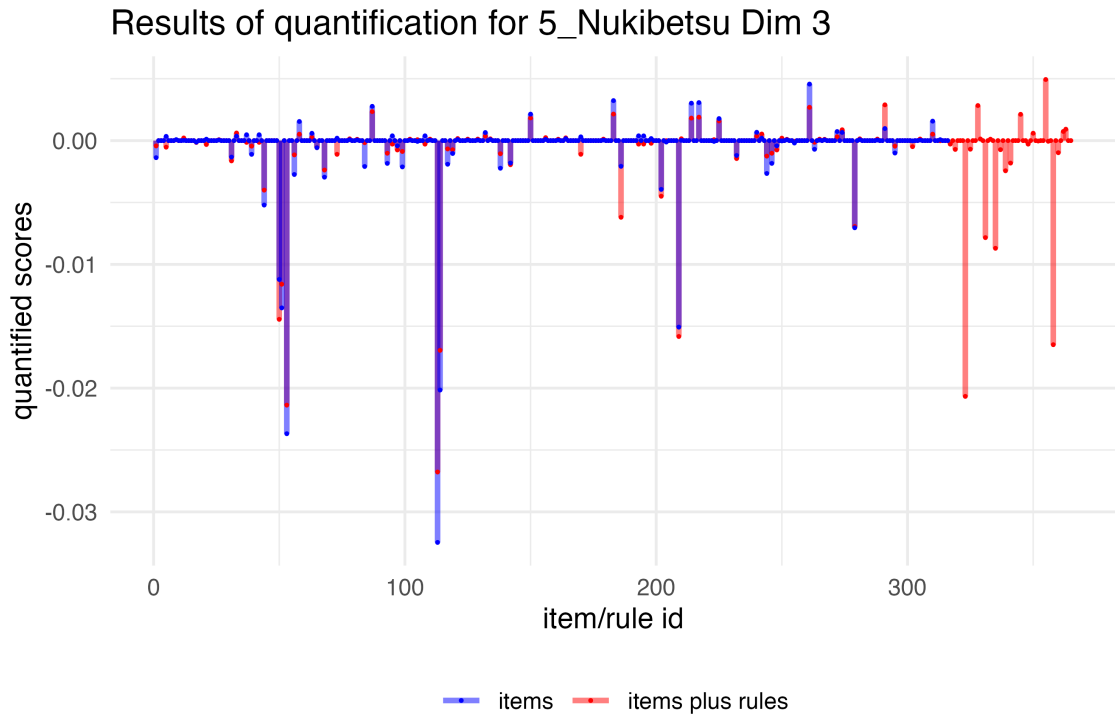


Figure 19. The quantified scores in the third dimension of the Nukibetsu dialect (No. 5).

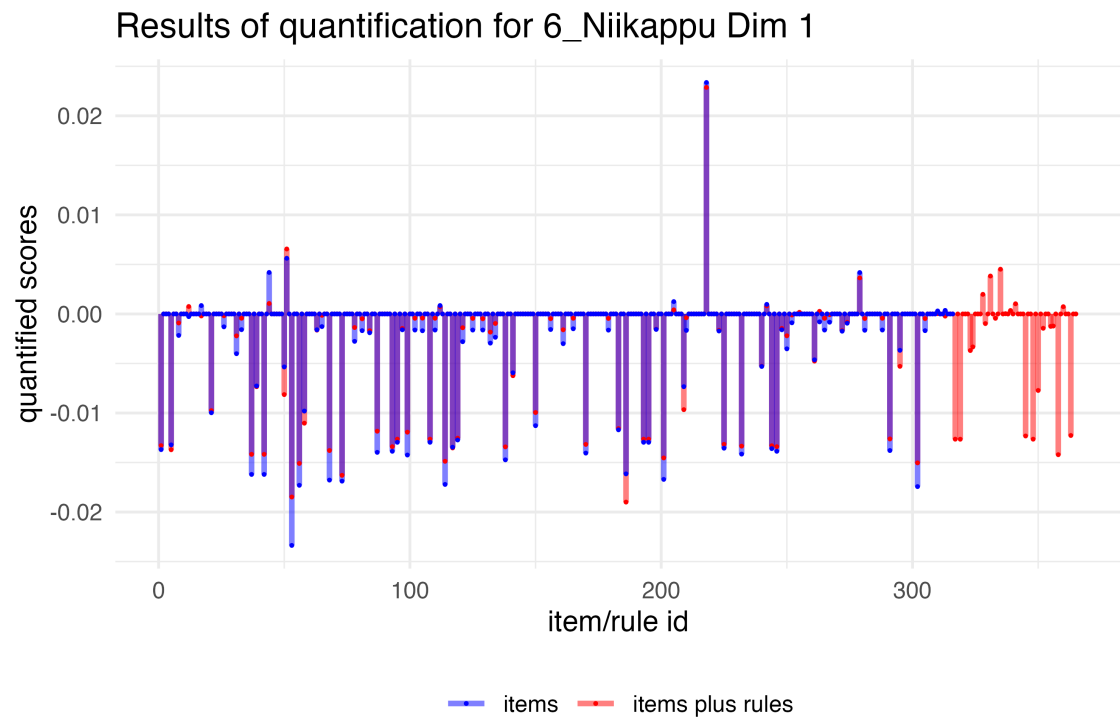


Figure 20. The quantified scores in the first dimension of the Niikappu dialect (No. 6 in Figure 1).

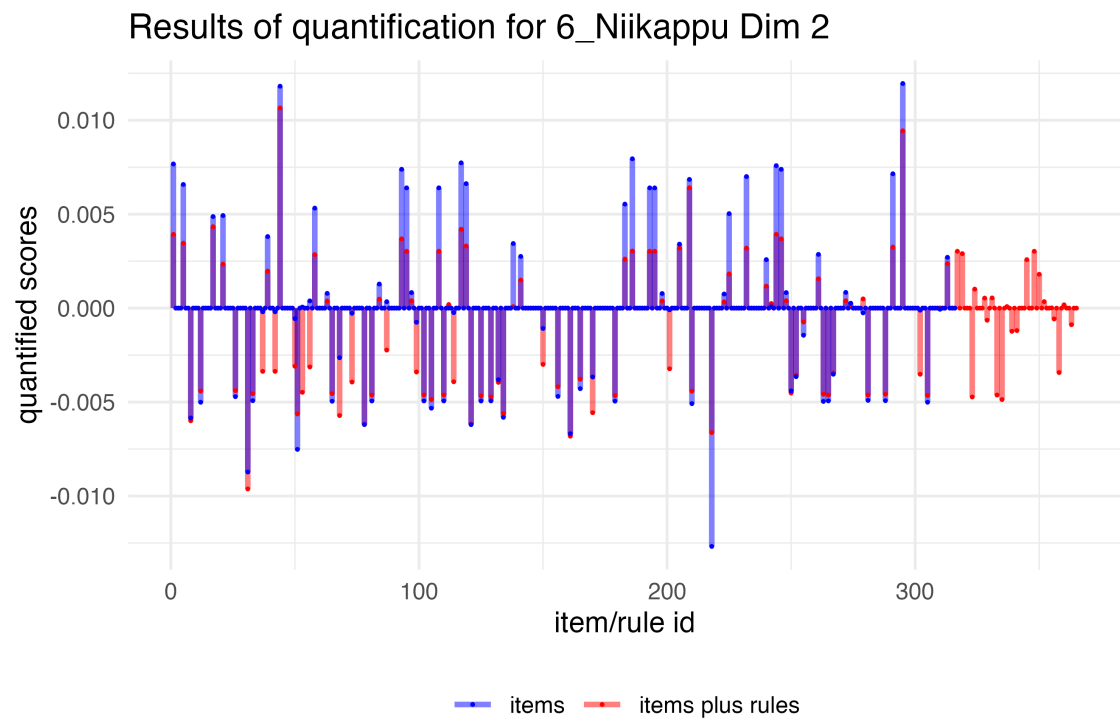


Figure 21. The quantified scores in the second dimension of the Niikappu dialect (No. 6).

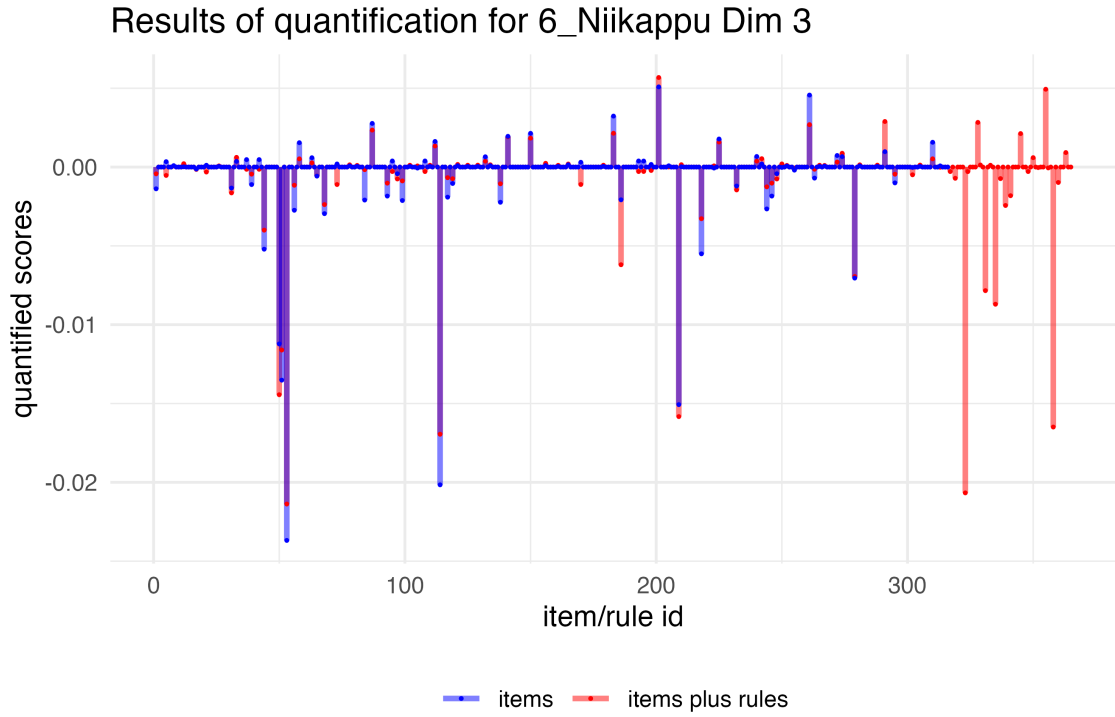


Figure 22. The quantified scores in the third dimension of the Niikappu dialect (No. 6).

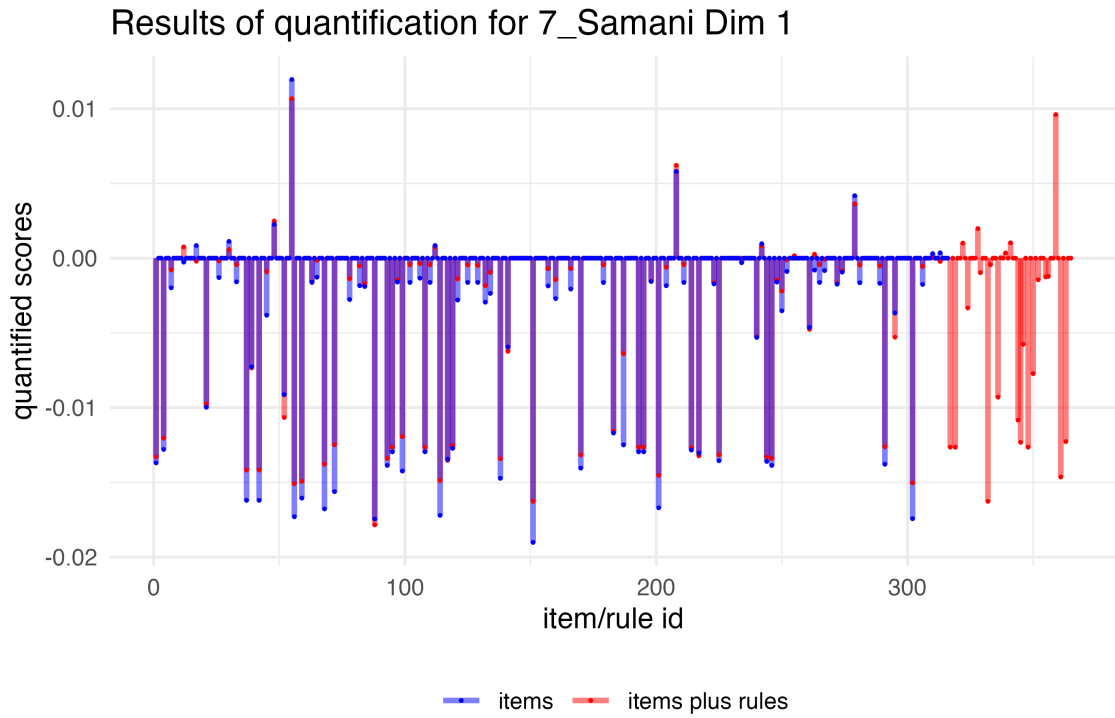


Figure 23. The quantified scores in the first dimension of the Samani dialect (No. 7 in Figure 1).

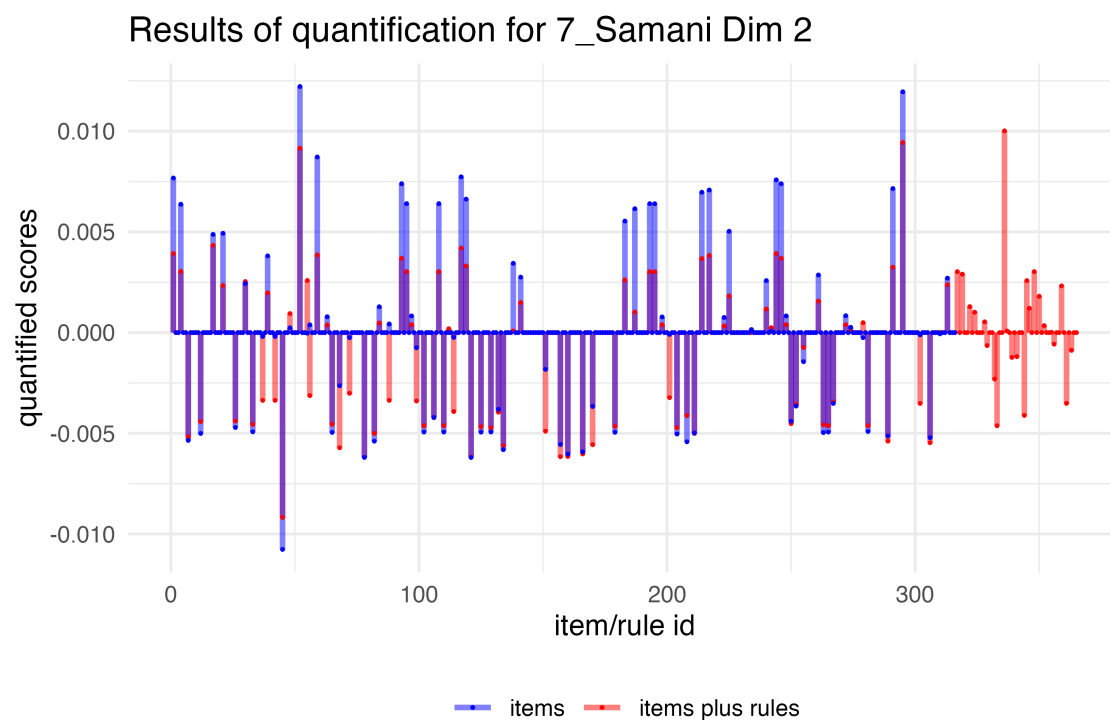


Figure 24. The quantified scores in the second dimension of the Samani dialect (No. 7).

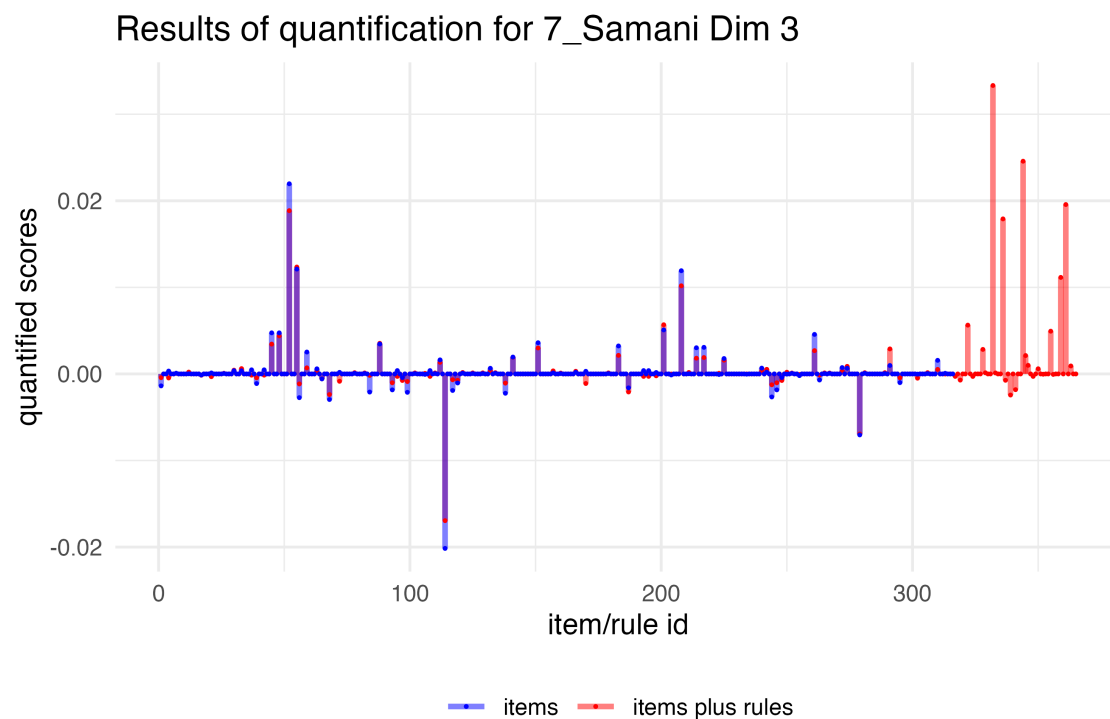


Figure 25. The quantified scores in the third dimension of the Samani dialect (No. 7).

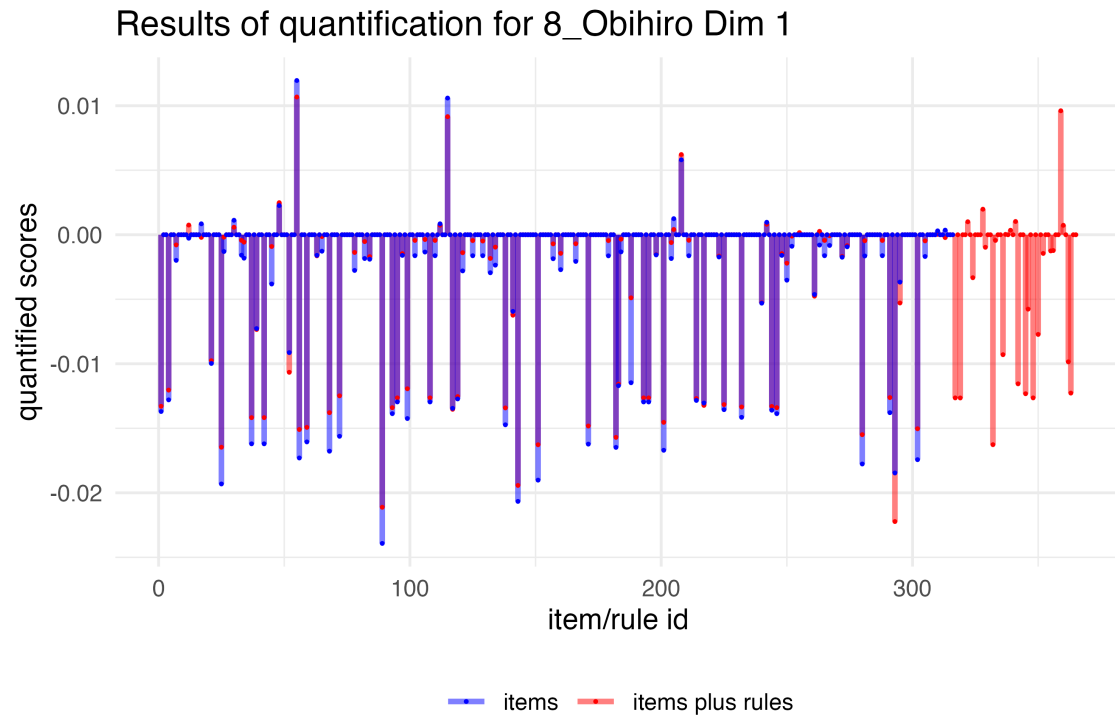


Figure 26. The quantified scores in the first dimension of the Obihiro dialect (No. 8 in Figure 1).

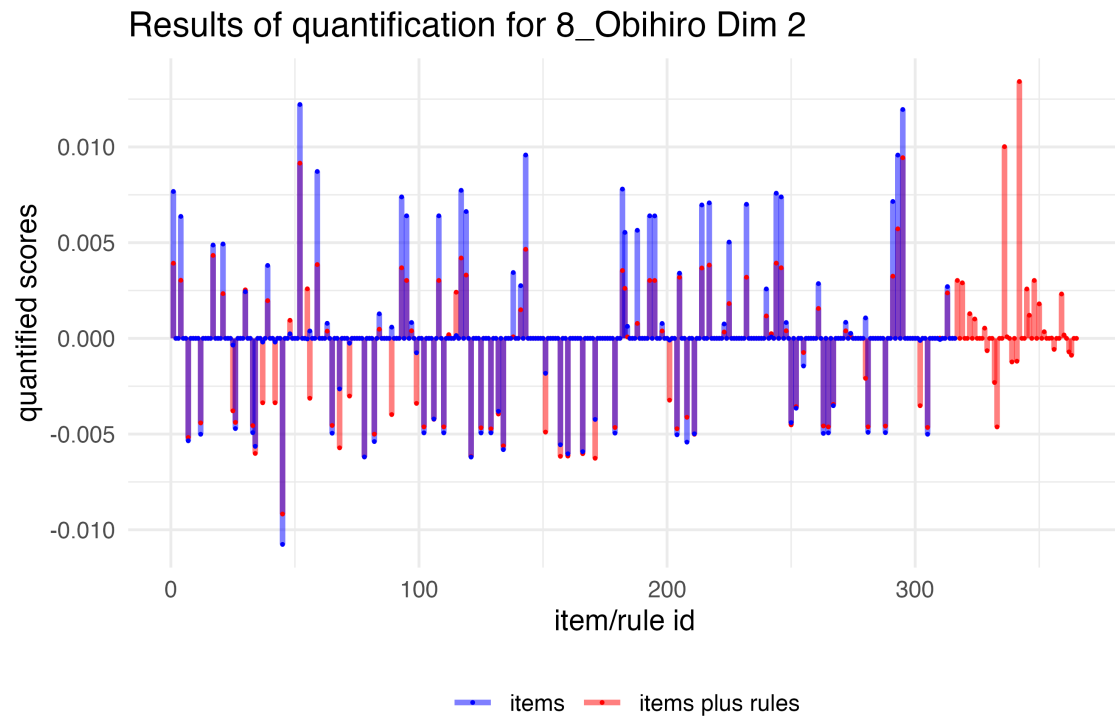


Figure 27. The quantified scores in the second dimension of the Obihiro dialect (No. 8).

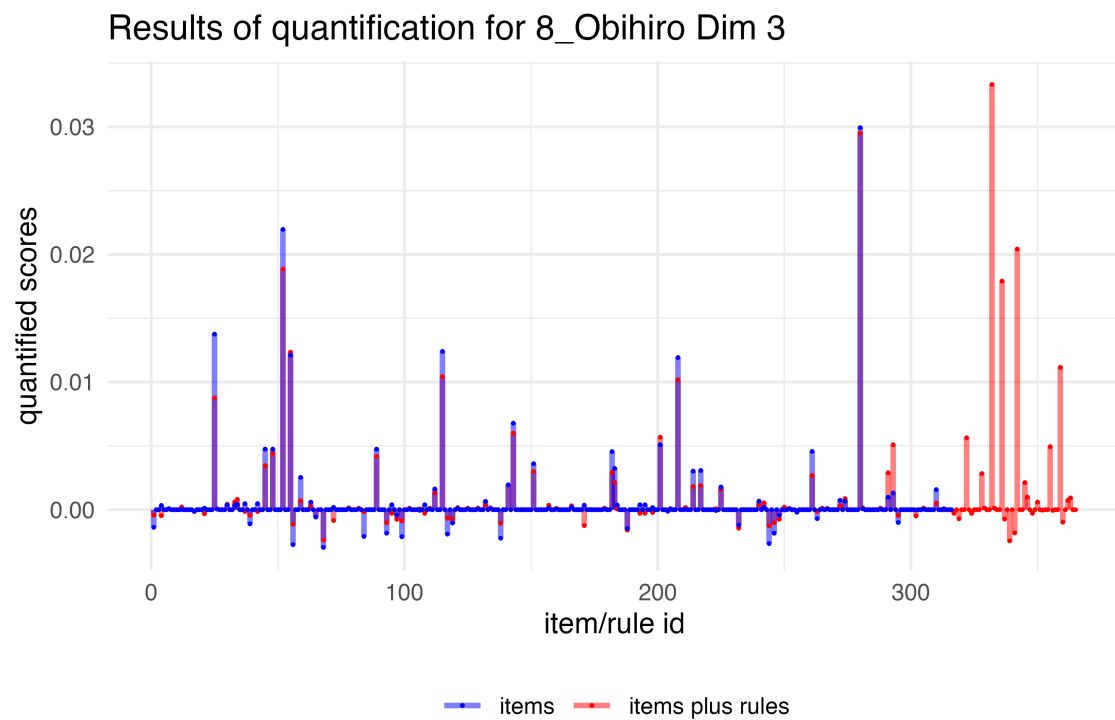


Figure 28. The quantified scores in the third dimension of the Obihiro dialect (No. 8).

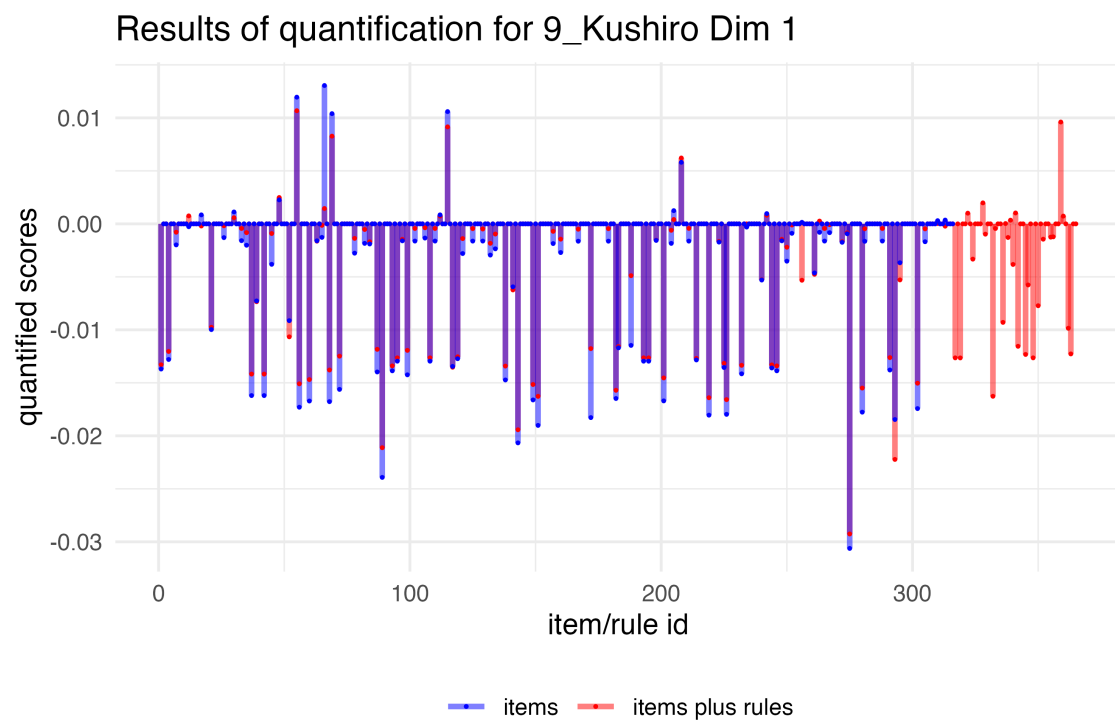


Figure 29. The quantified scores in the first dimension of the Kushiro dialect (No. 9 in Figure 1).

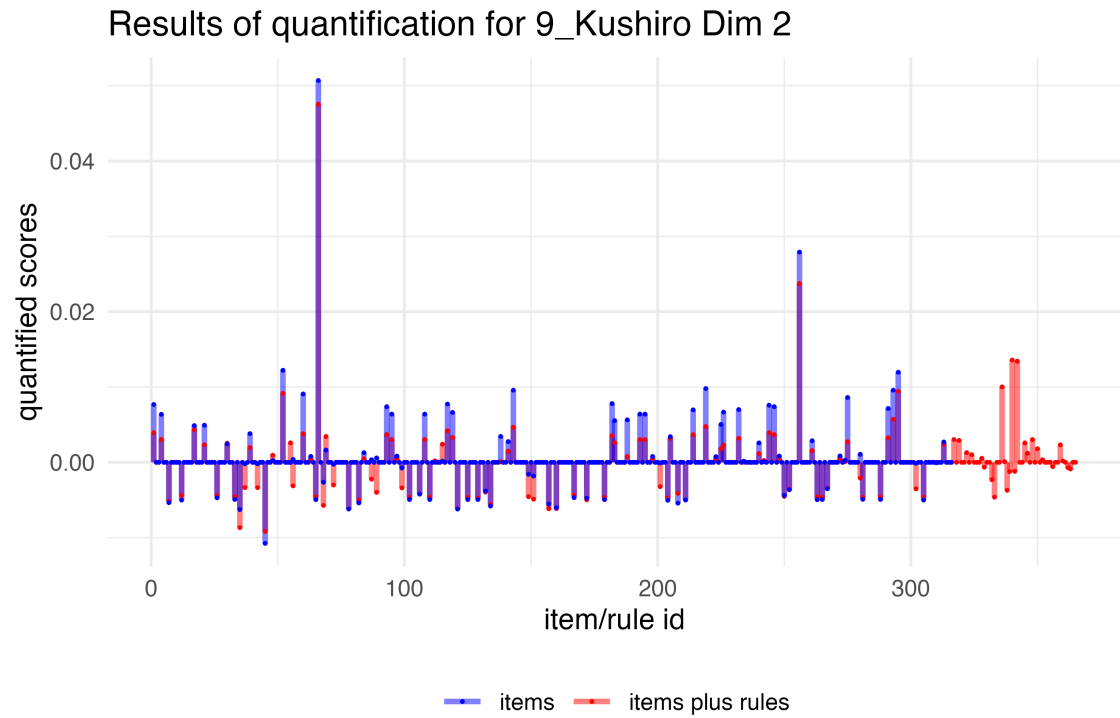


Figure 30. The quantified scores in the second dimension of the Kushiro dialect (No. 9).

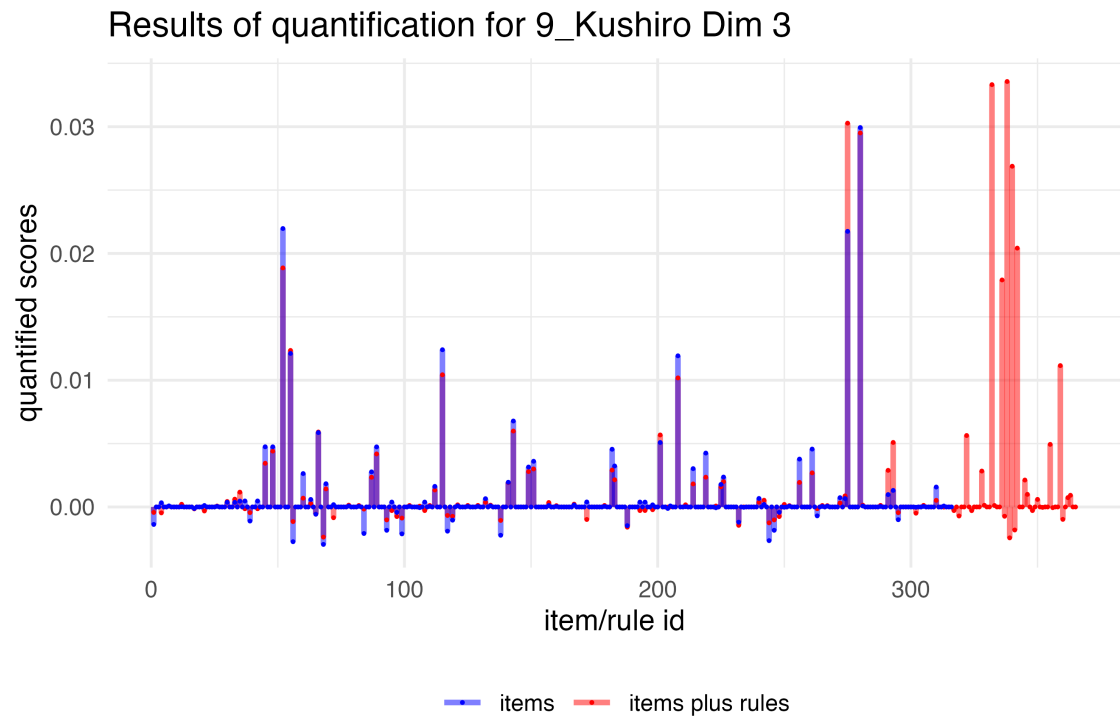


Figure 31. The quantified scores in the third dimension of the Kushiro dialect (No. 9).

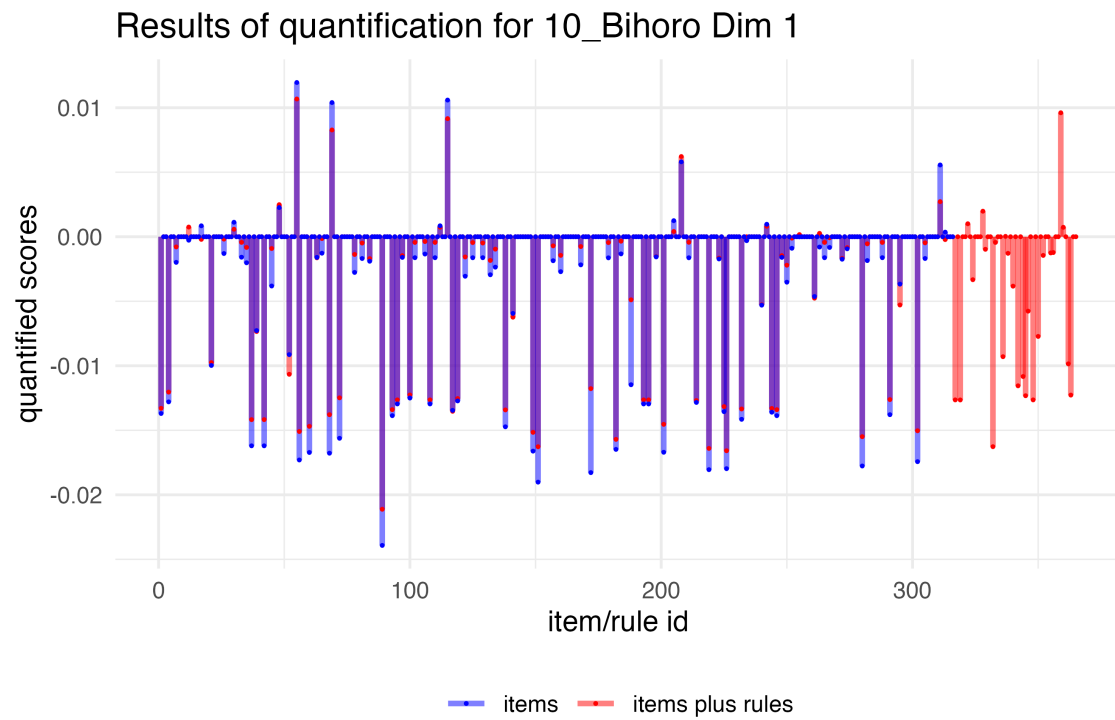


Figure 32. The quantified scores in the first dimension of the Bihoro dialect (No. 10 in Figure 1).

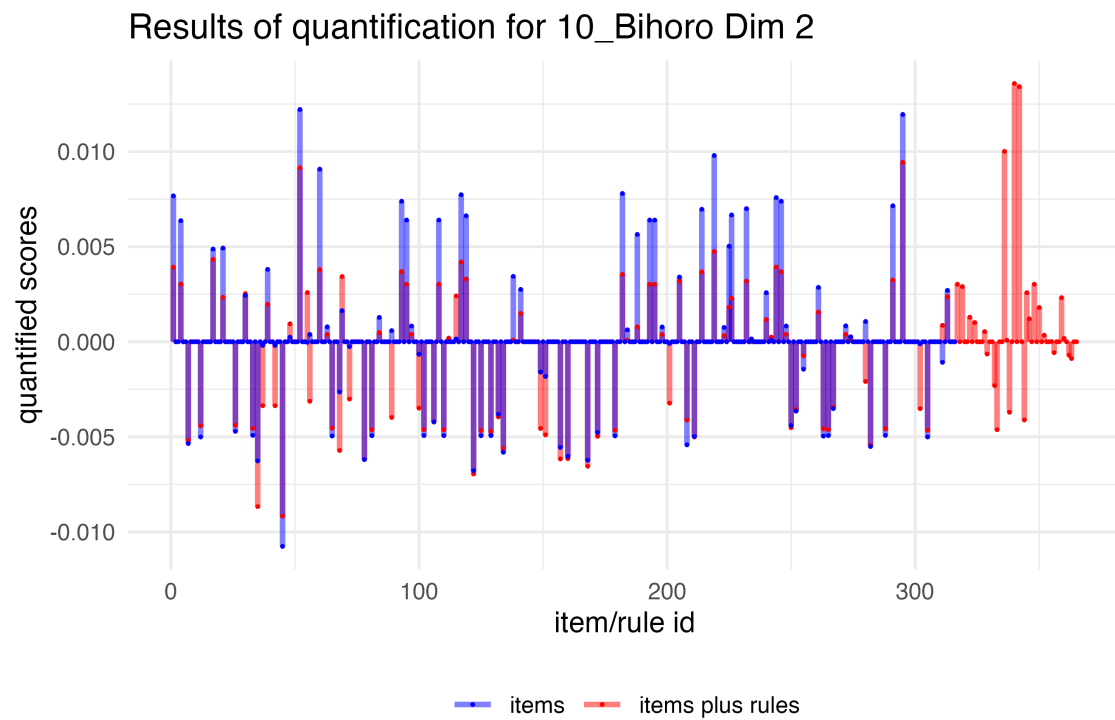


Figure 33. The quantified scores in the second dimension of the Bihoro dialect (No. 10).

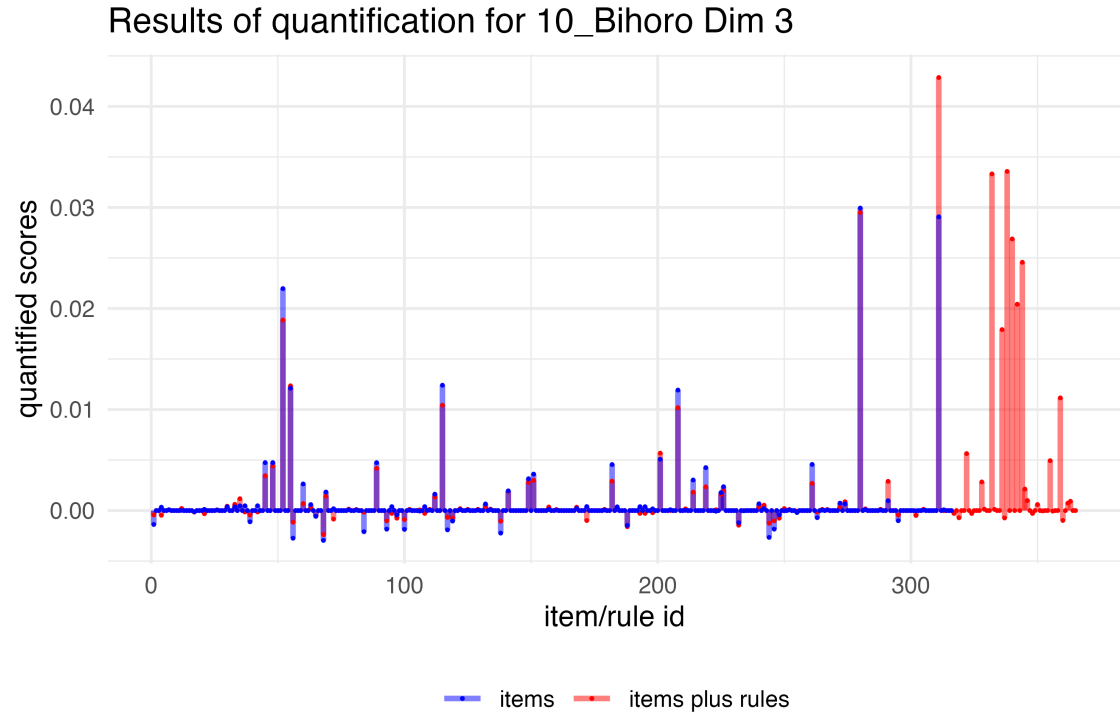


Figure 34. The quantified scores in the third dimension of the Bihoro dialect (No. 10).

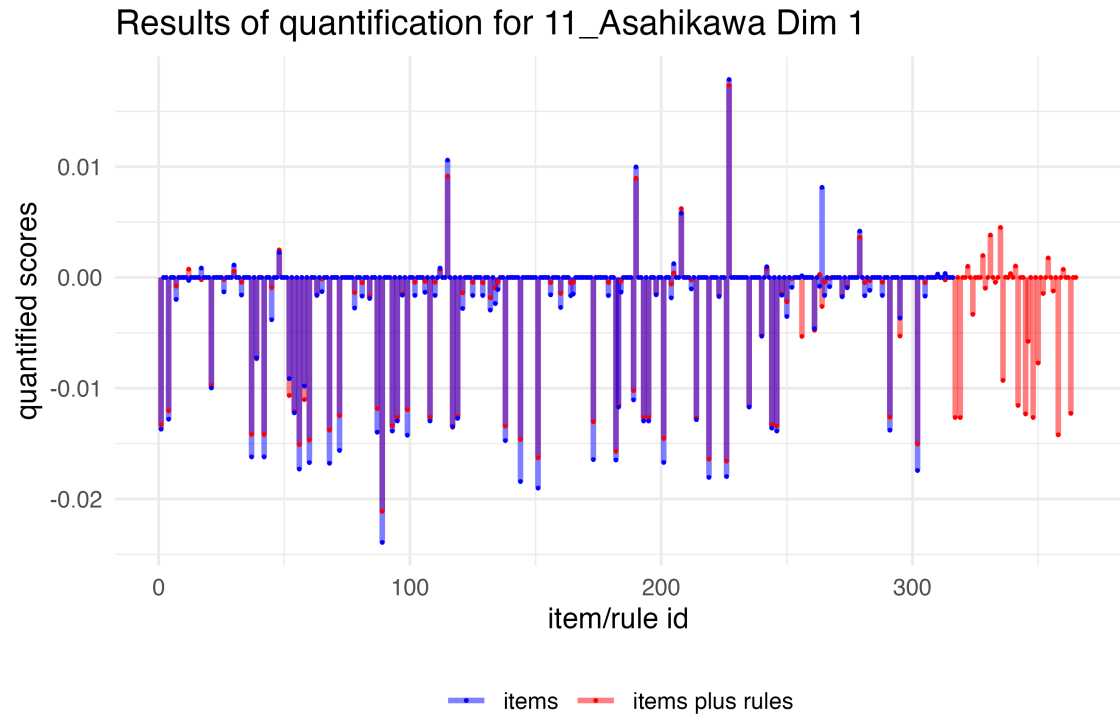


Figure 35. The quantified scores in the first dimension of the Ashikawa dialect (No. 11 in Figure 1).

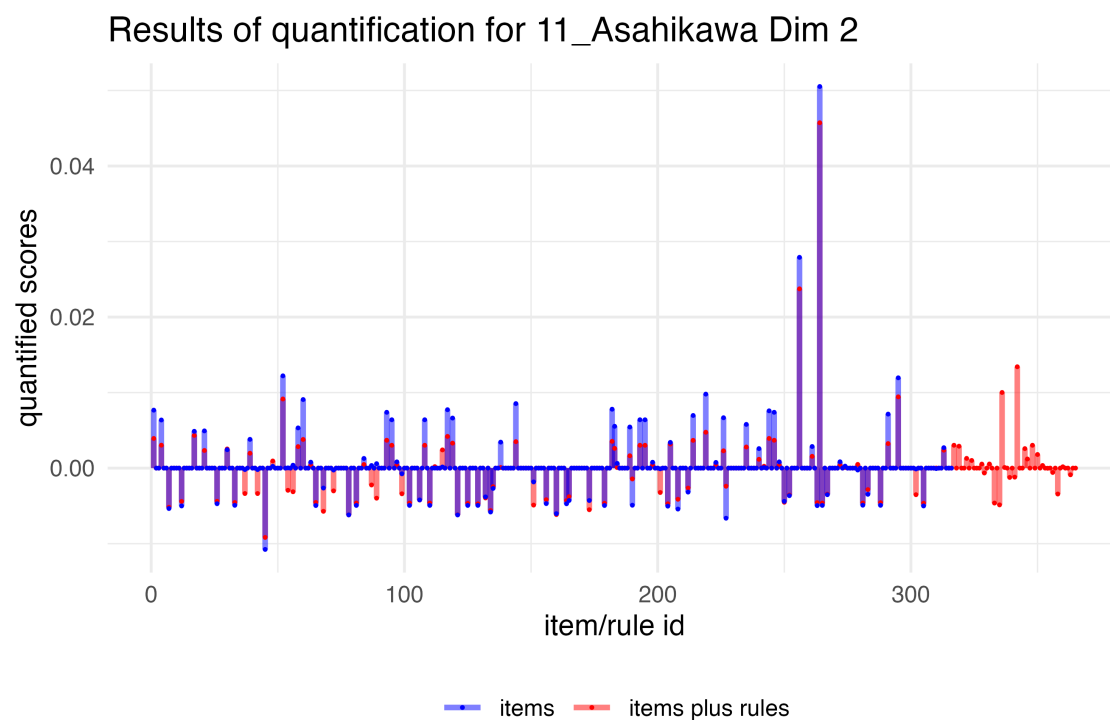


Figure 36. The quantified scores in the second dimension of the Asahikawa dialect (No. 11).

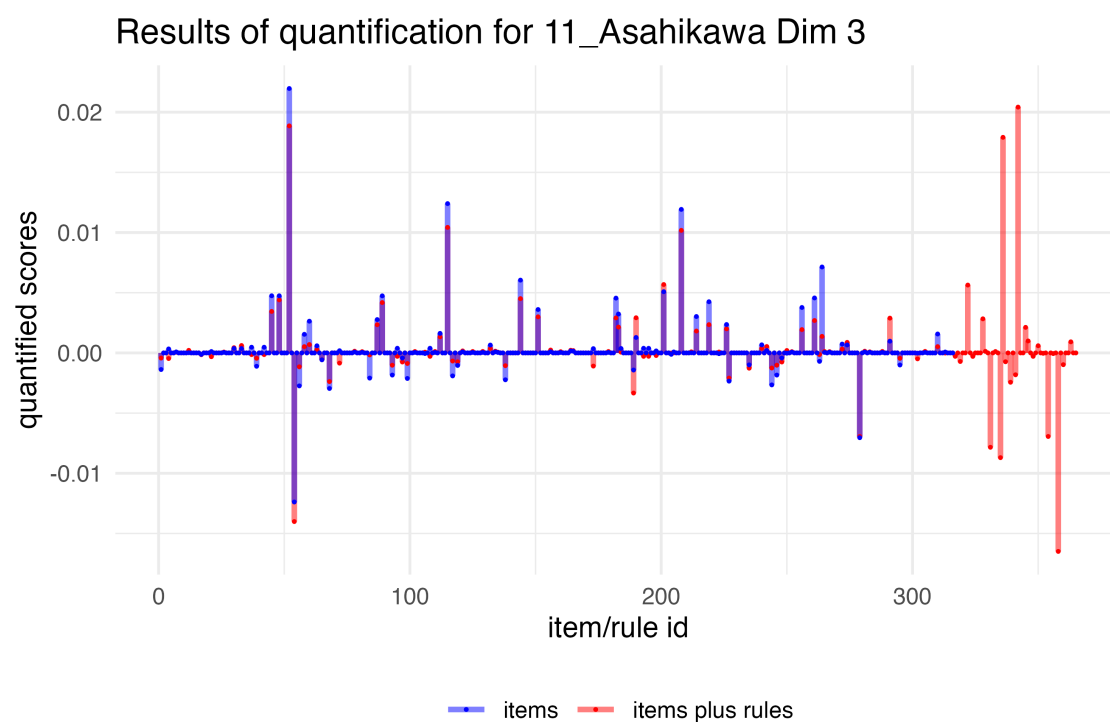


Figure 37. The quantified scores in the third dimension of the Asahikawa dialect (No. 11).

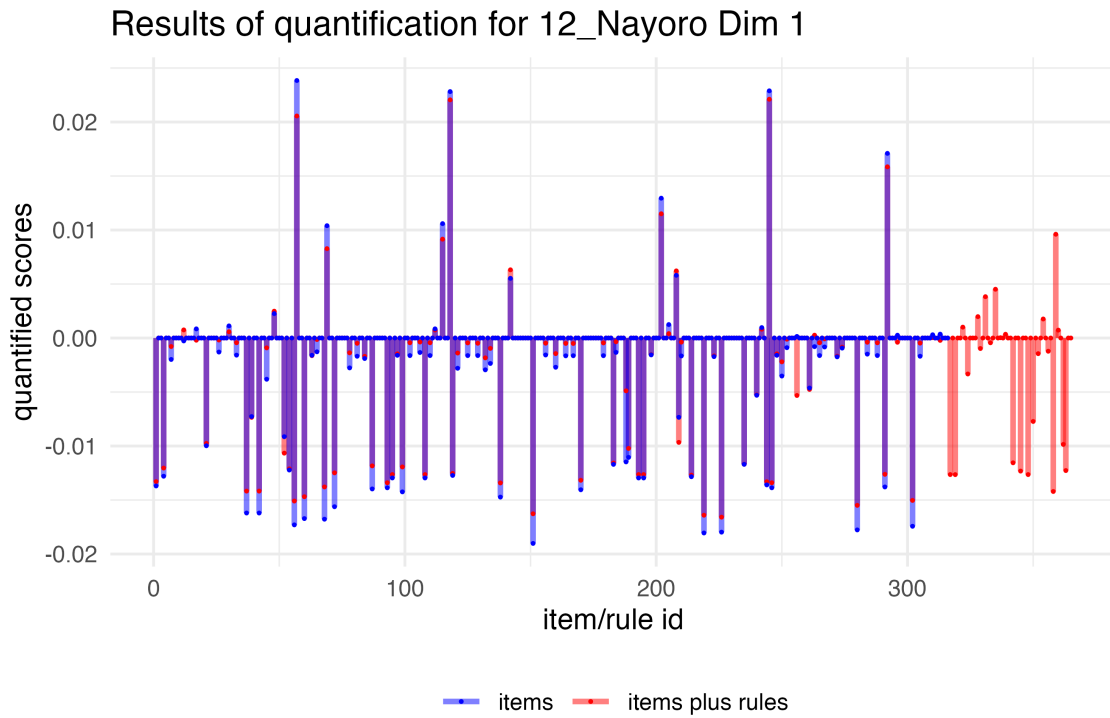


Figure 38. The quantified scores in the first dimension of the Nayoro dialect (No. 12 in Figure 1).

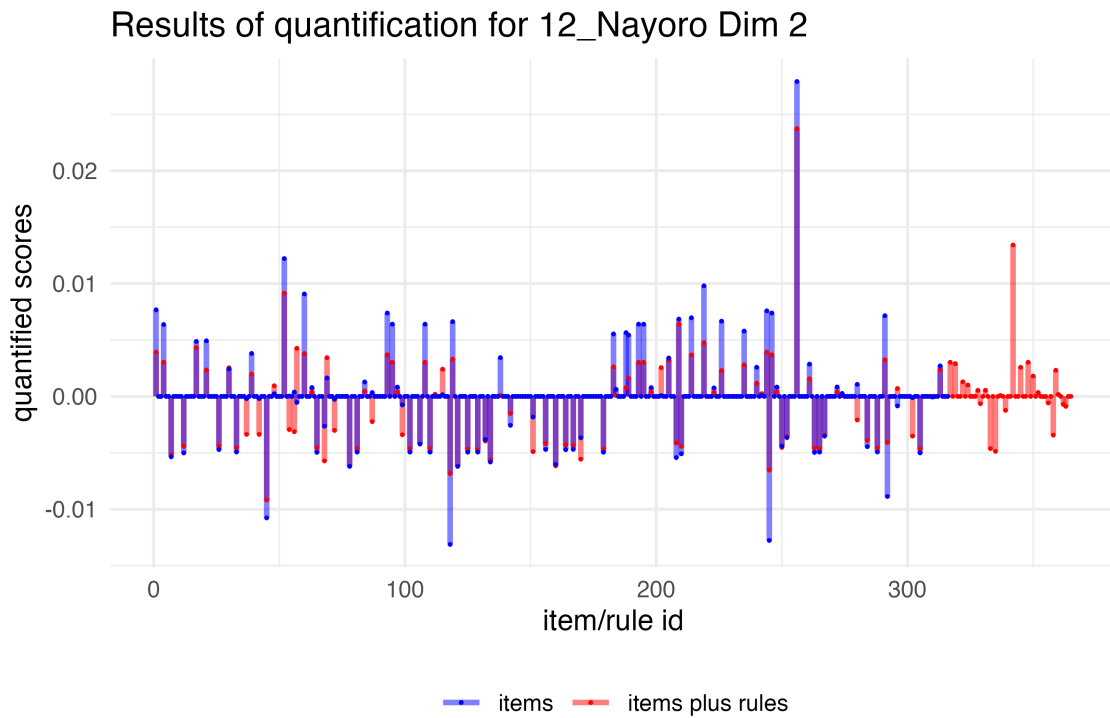


Figure 39. The quantified scores in the second dimension of the Nayoro dialect (No. 12).

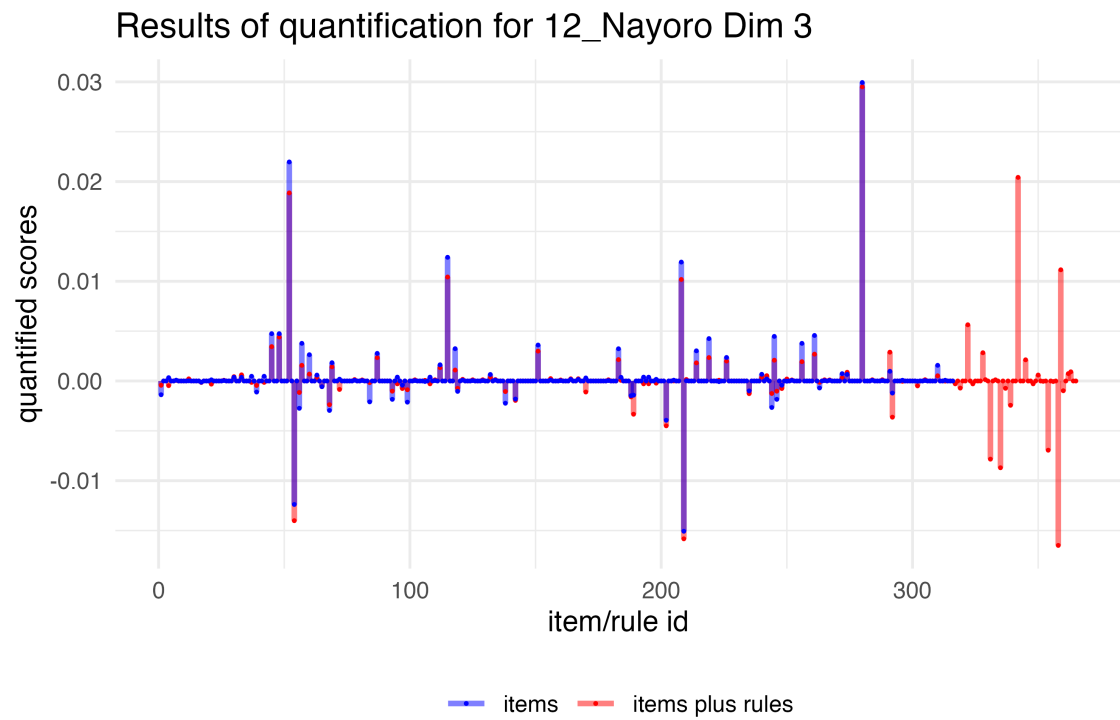


Figure 40. The quantified scores in the third dimension of the Nayoro dialect (No. 12).

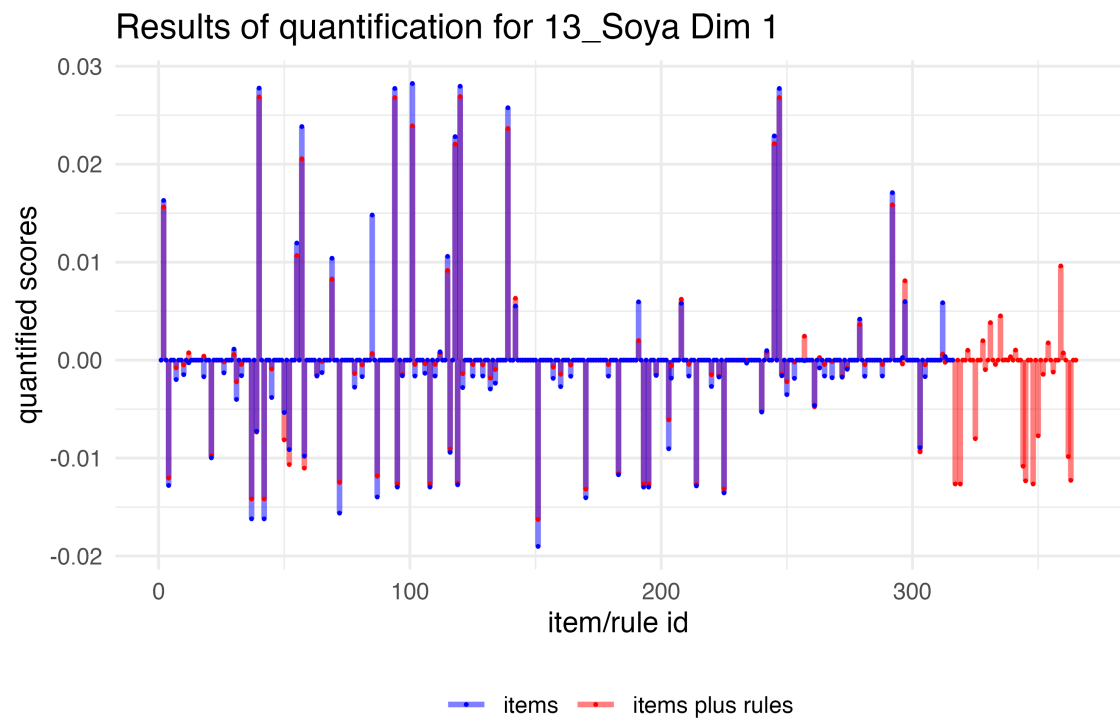


Figure 41. The quantified scores in the first dimension of the Soya dialect (No. 13 in Figure 1).

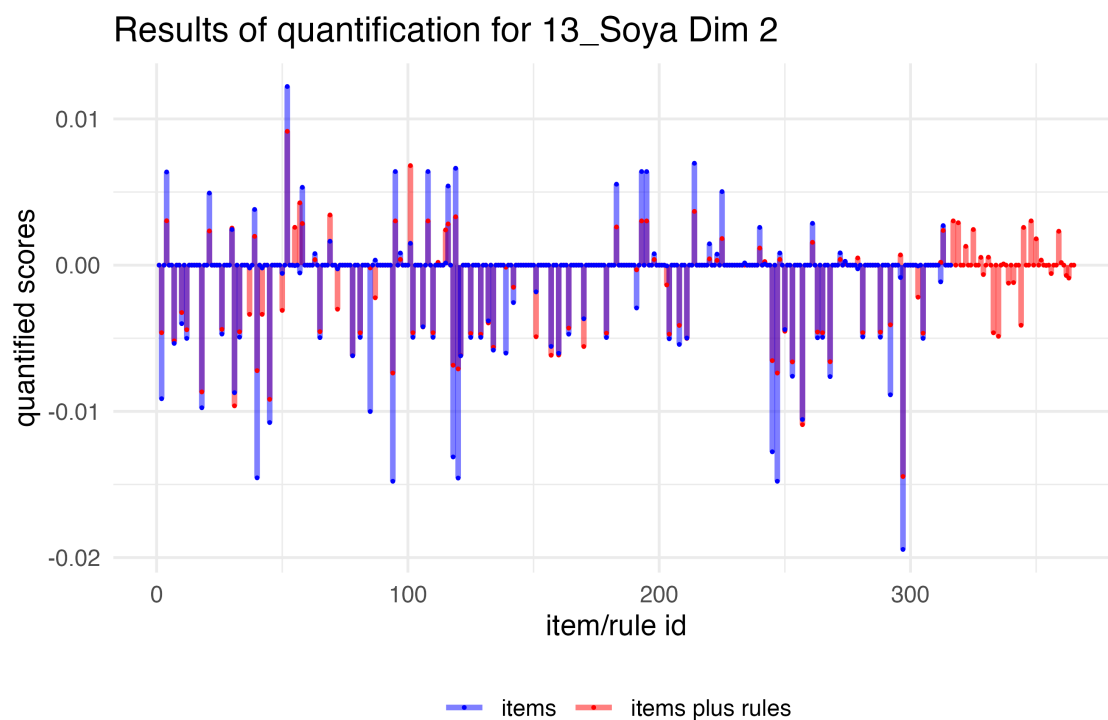


Figure 42. The quantified scores in the second dimension of the Soya dialect (No. 13).

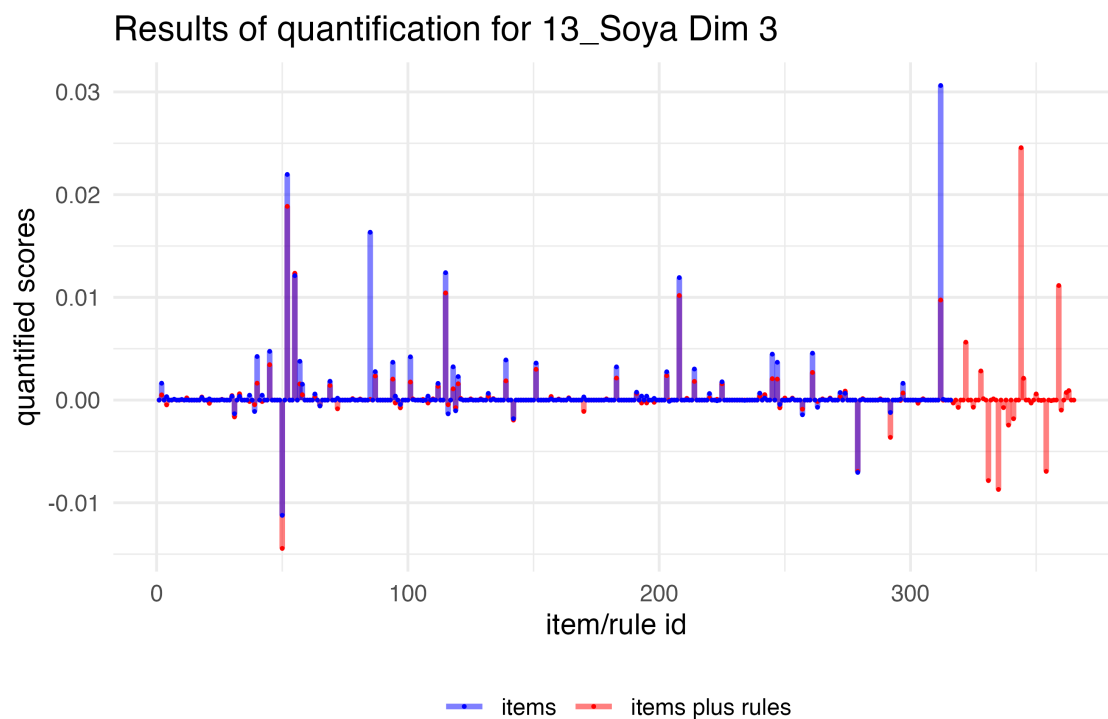


Figure 43. The quantified scores in the third dimension of the Soya dialect (No. 13).

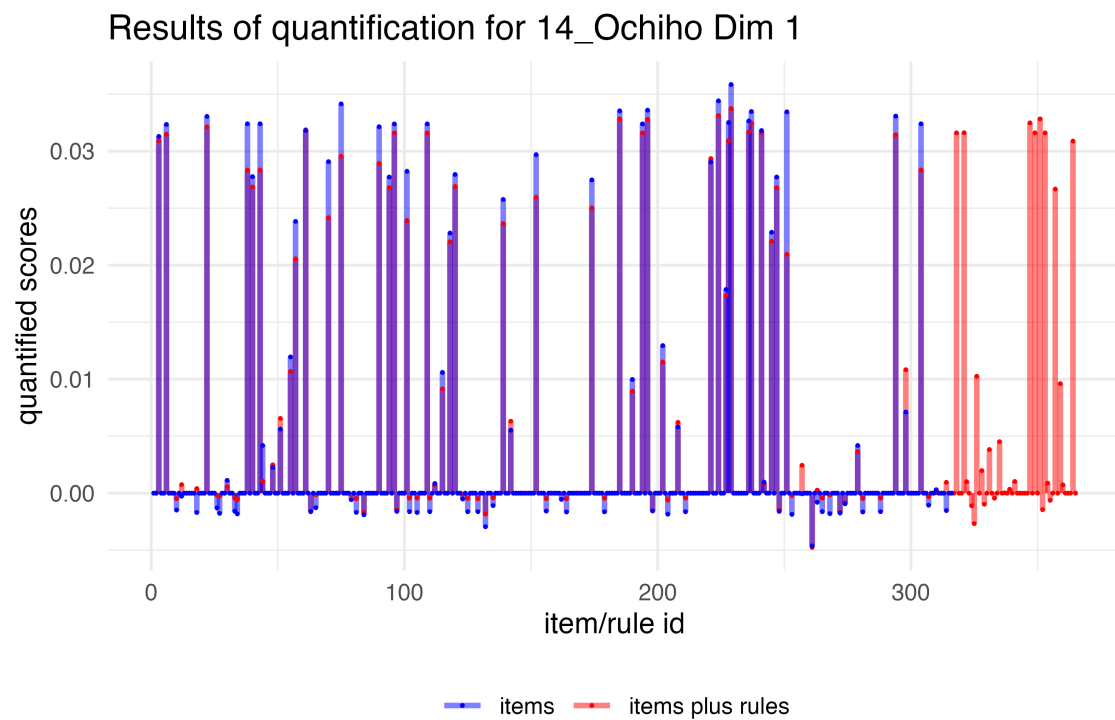


Figure 44. The quantified scores in the first dimension of the Ochiho dialect (No. 14 in Figure 1).

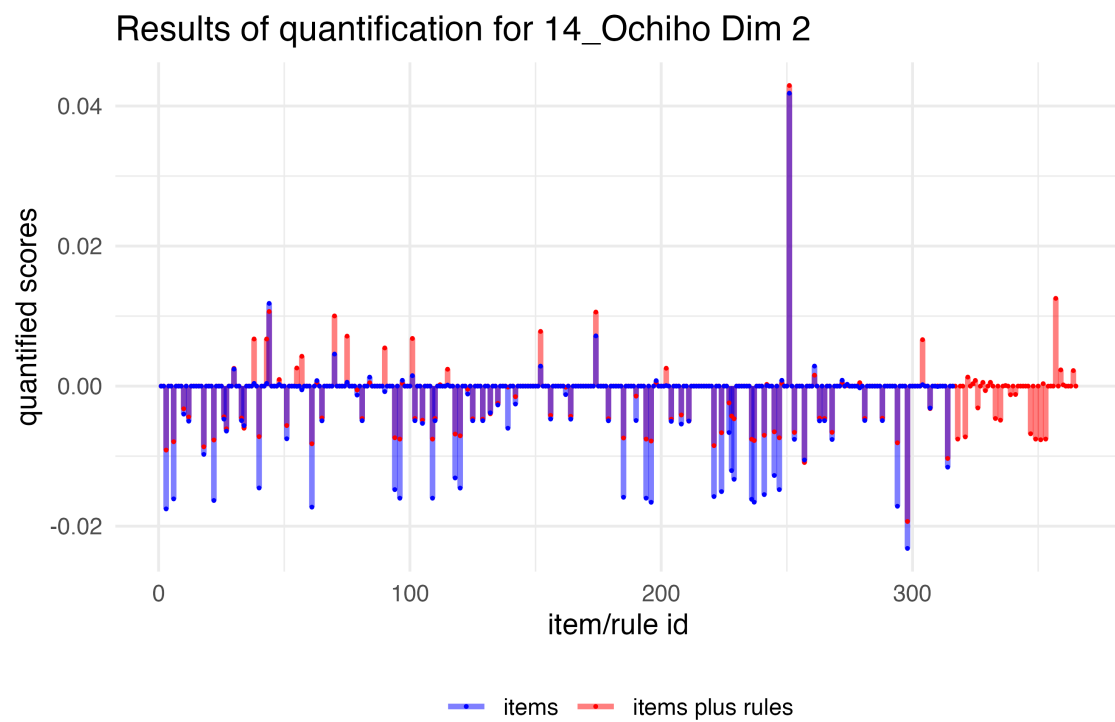


Figure 45. The quantified scores in the second dimension of the Ochiho dialect (No. 14).

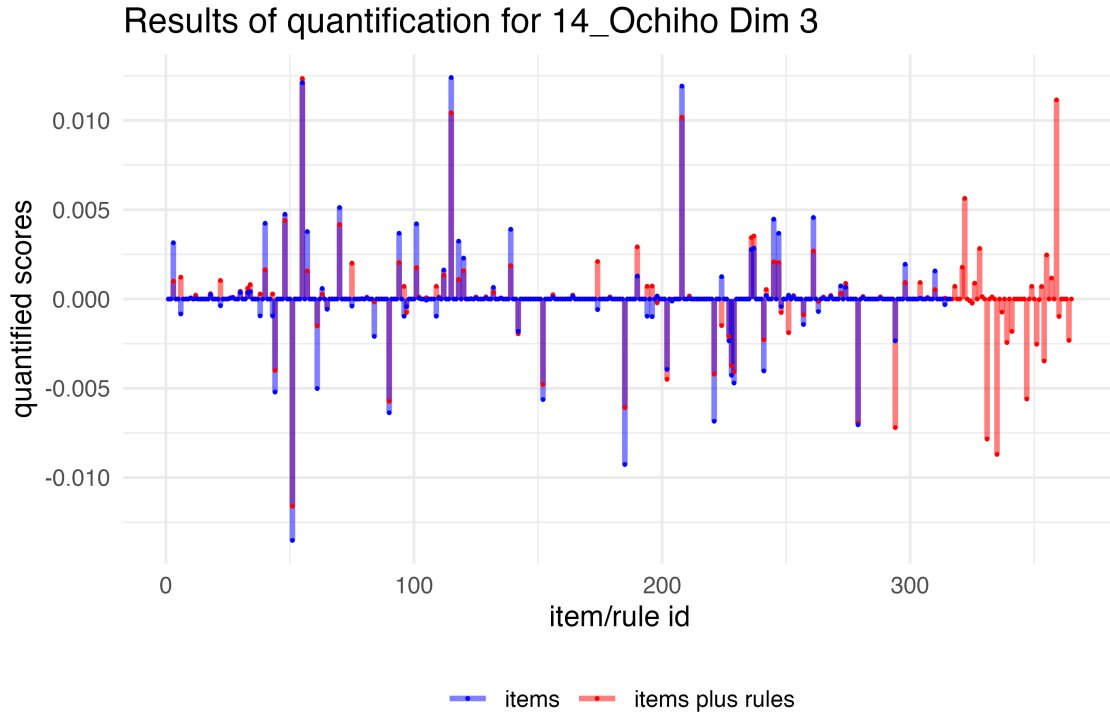


Figure 46. The quantified scores in the third dimension of the Ochiho dialect (No. 14).

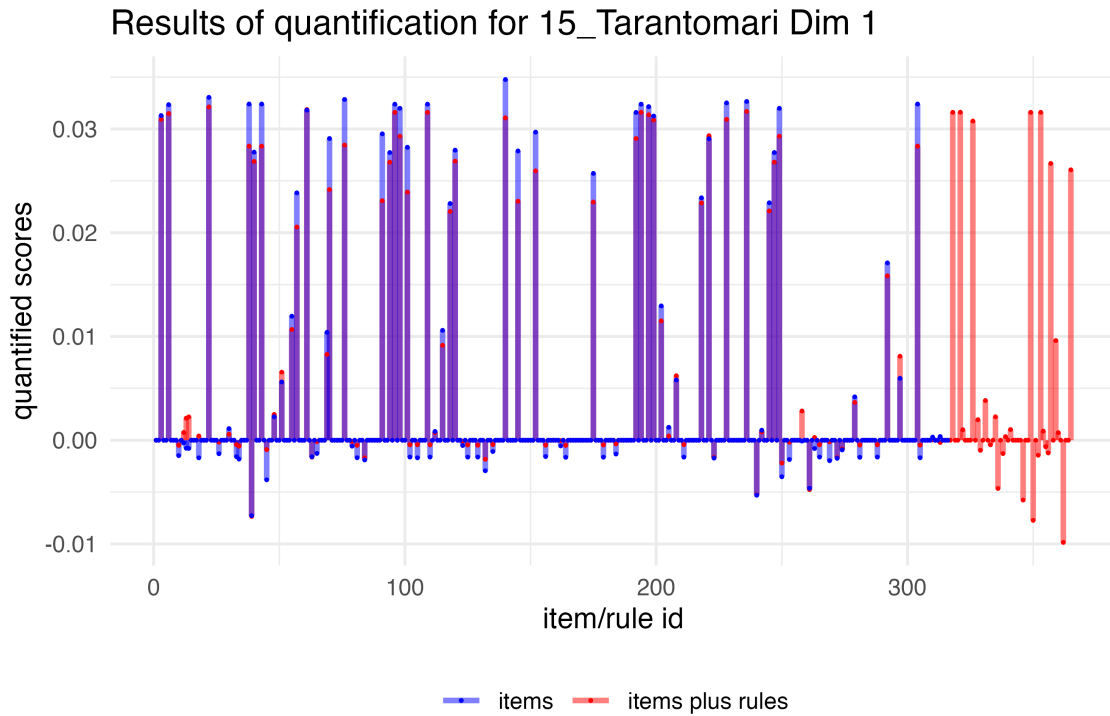


Figure 47. The quantified scores in the first dimension of the Tarantomari dialect (No. 15 in Figure 1).

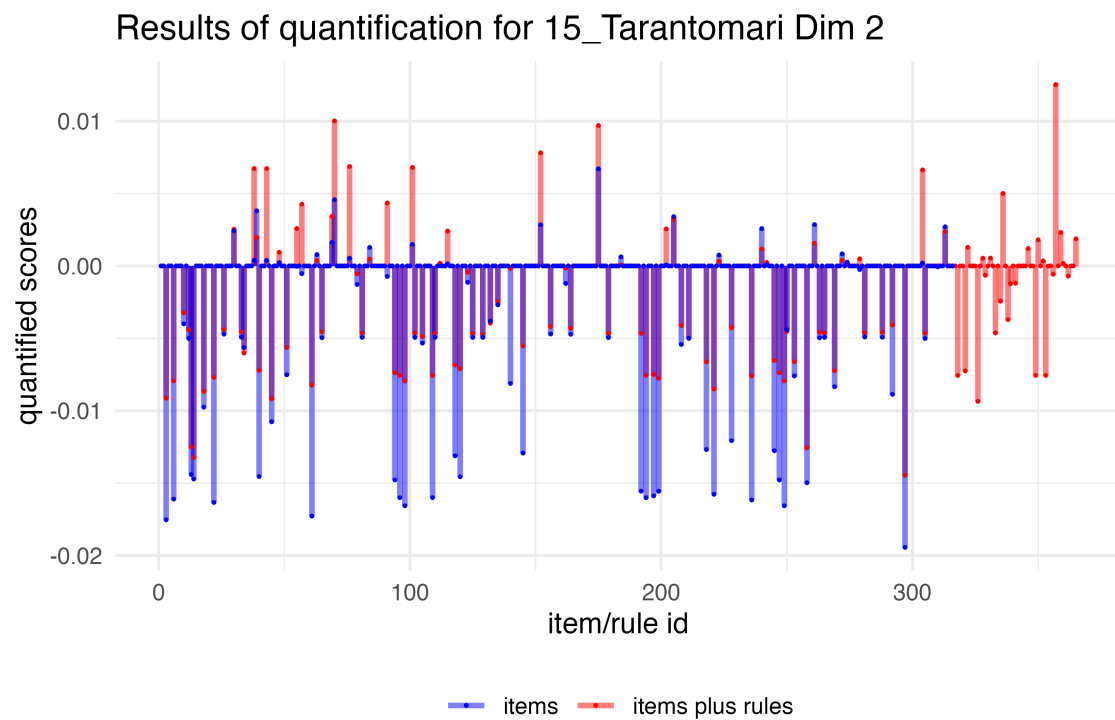


Figure 48. The quantified scores in the second dimension of the Tarantomari dialect (No. 15).

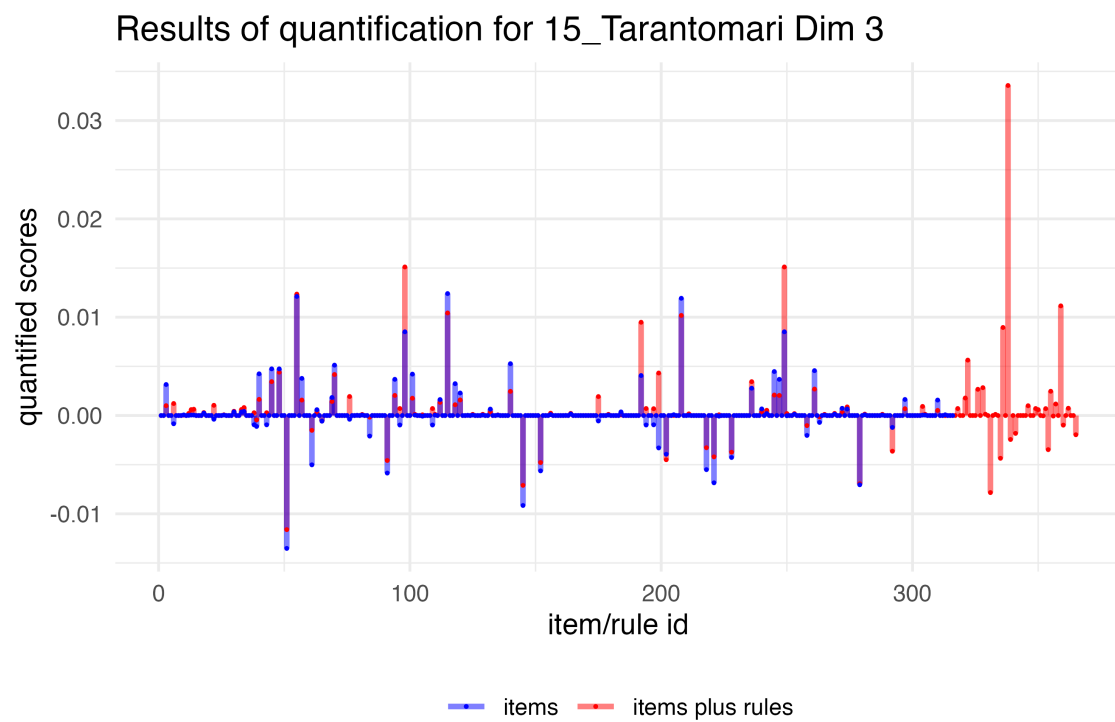


Figure 49. The quantified scores in the third dimension of the Tarantomari dialect (No. 15).

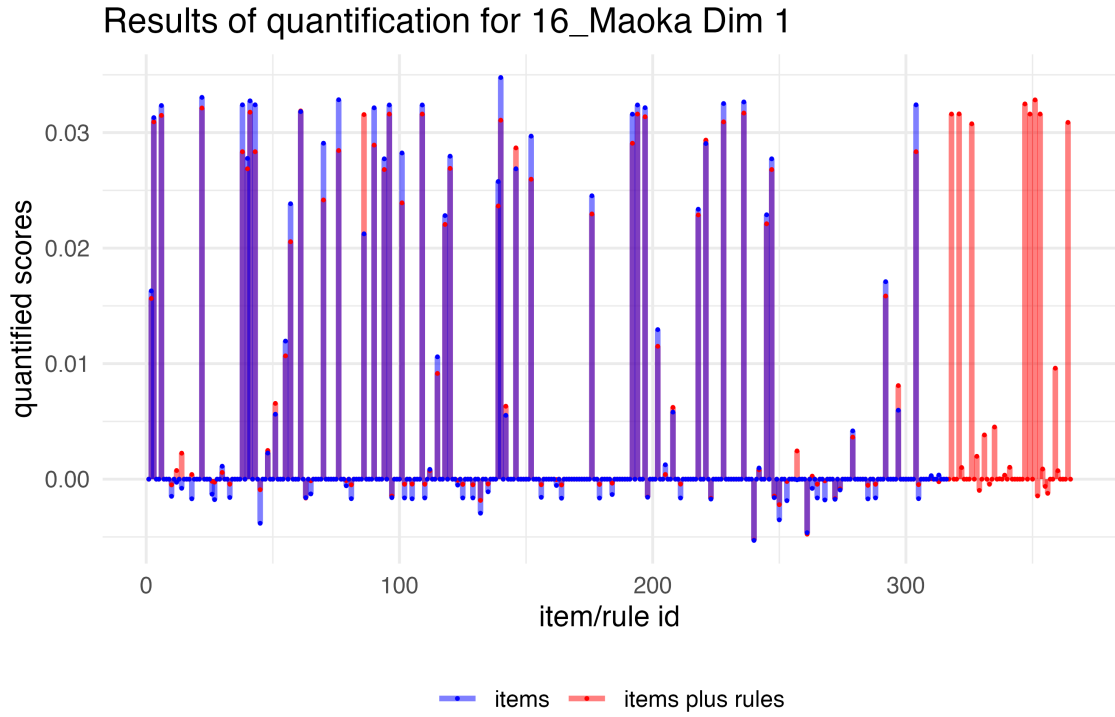


Figure 50. The quantified scores in the first dimension of the Maoka dialect (No. 16 in Figure 1).

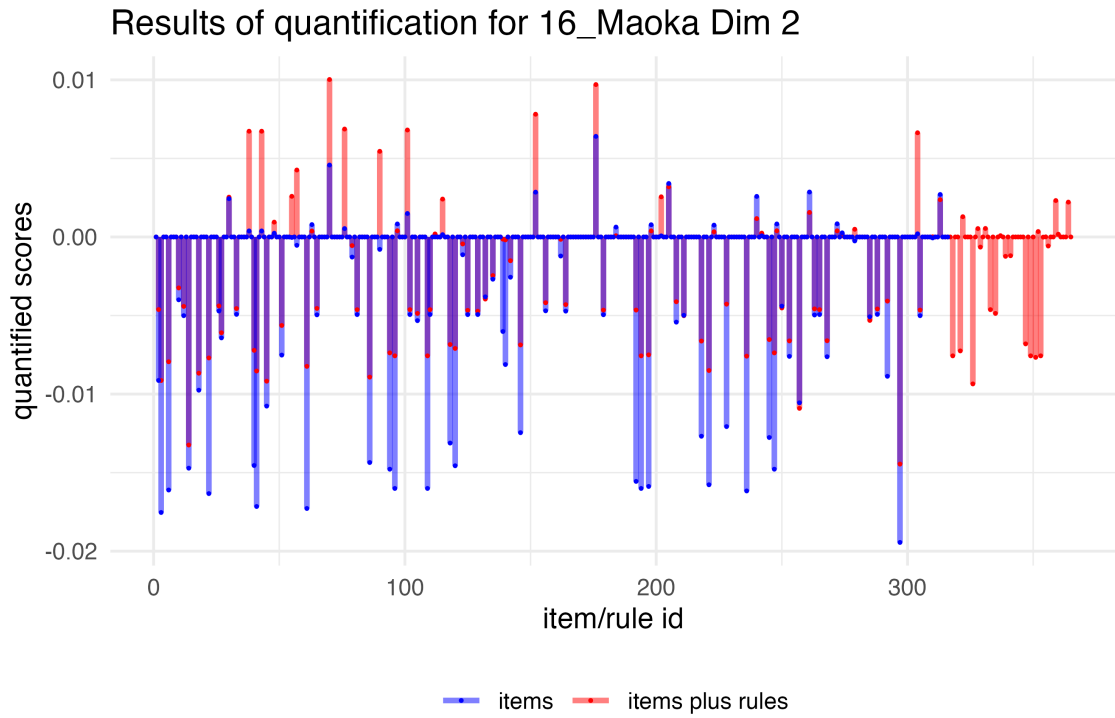


Figure 51. The quantified scores in the second dimension of the Maoka dialect (No. 16).

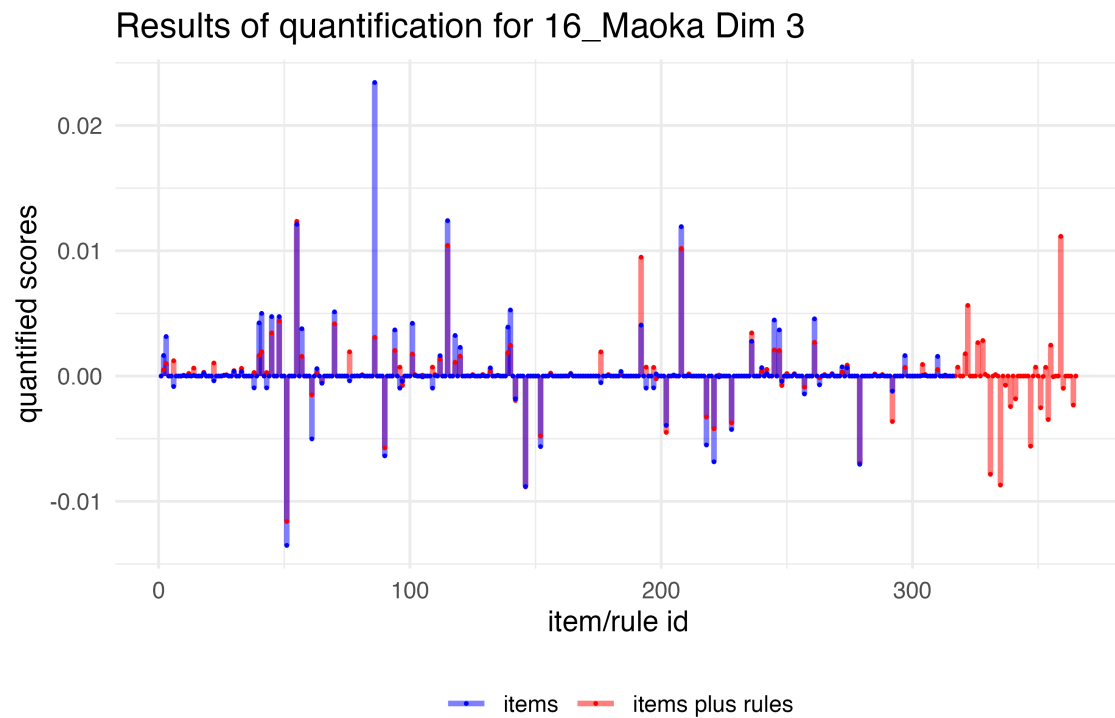


Figure 52. The quantified scores in the third dimension of the Maoka dialect (No. 16).

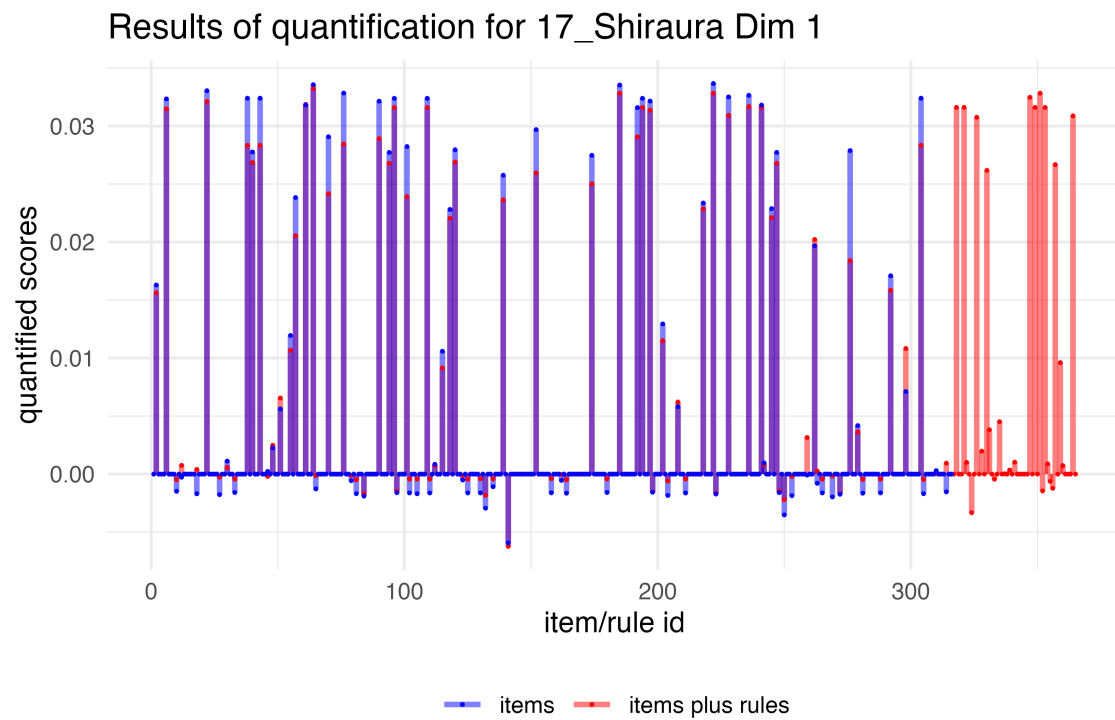


Figure 53. The quantified scores in the first dimension of the Shiraure dialect (No. 17 in Figure 1).

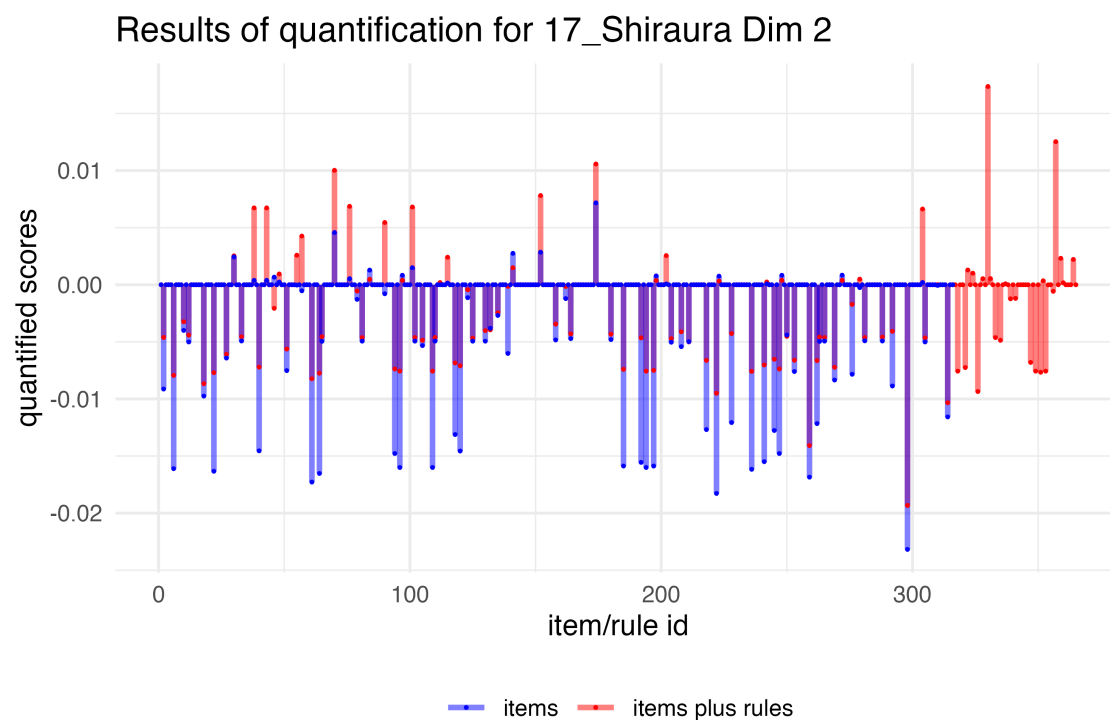


Figure 54. The quantified scores in the second dimension of the Shiraura dialect (No. 17).

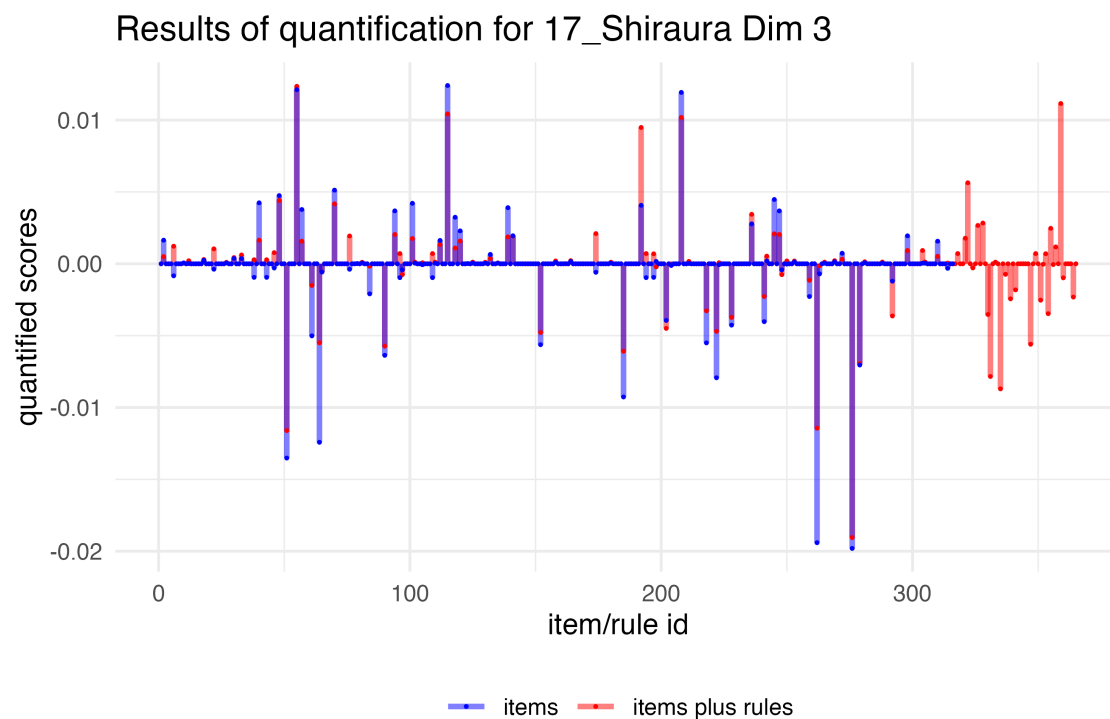


Figure 55. The quantified scores in the third dimension of the Shiraura dialect (No. 17).

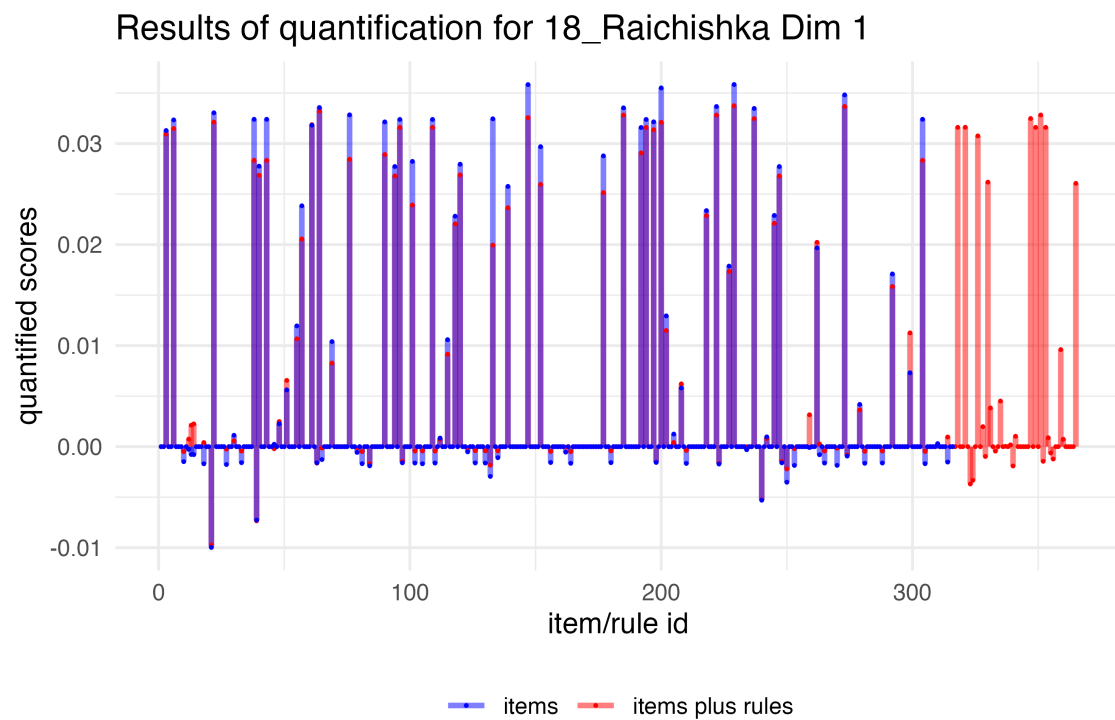


Figure 56. The quantified scores in the first dimension of the Raichishka dialect (No. 18 in Figure 1).

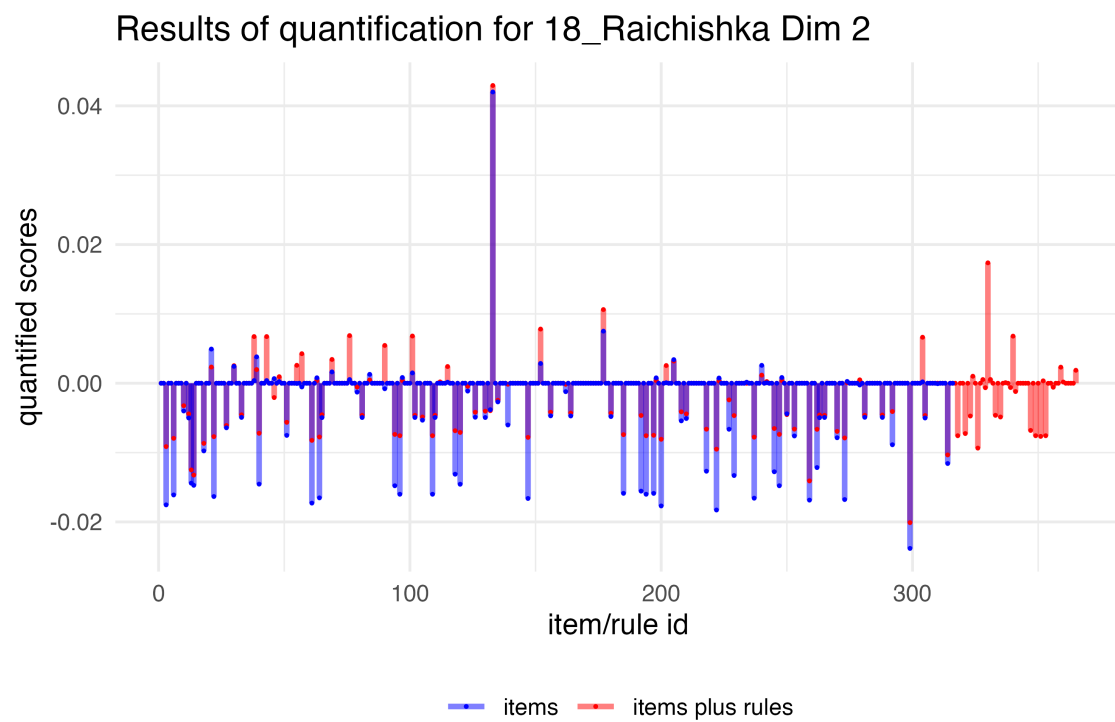


Figure 57. The quantified scores in the second dimension of the Raichishka dialect (No. 18).

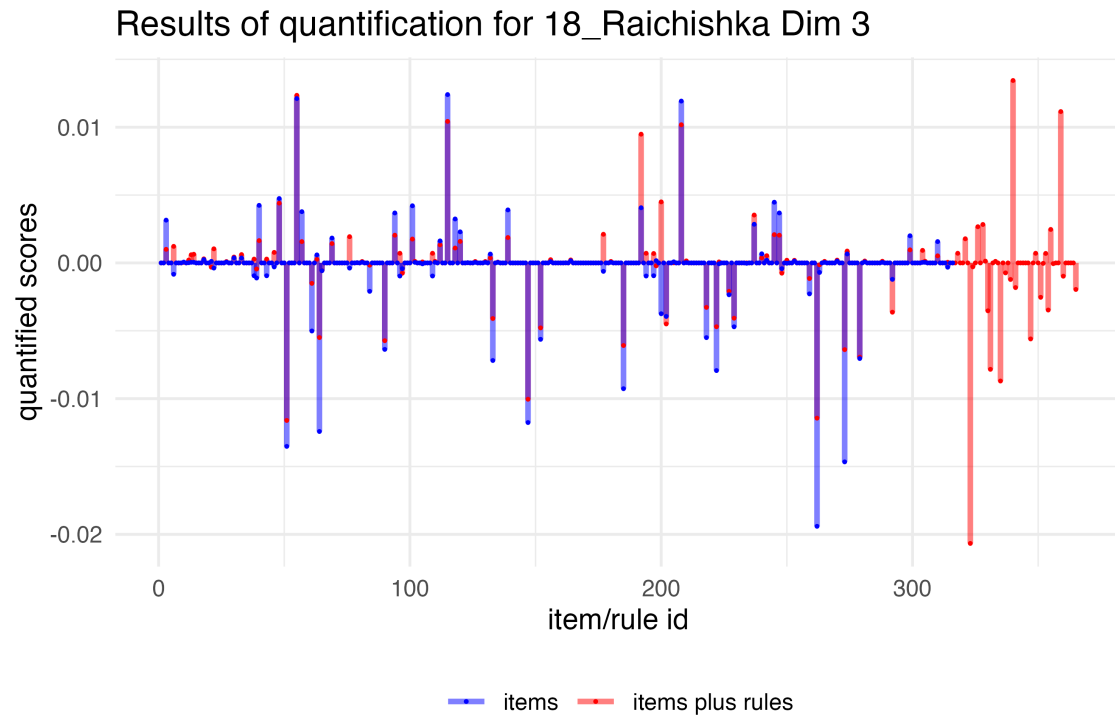


Figure 58. The quantified scores in the third dimension of the Raichishka dialect (No. 18).

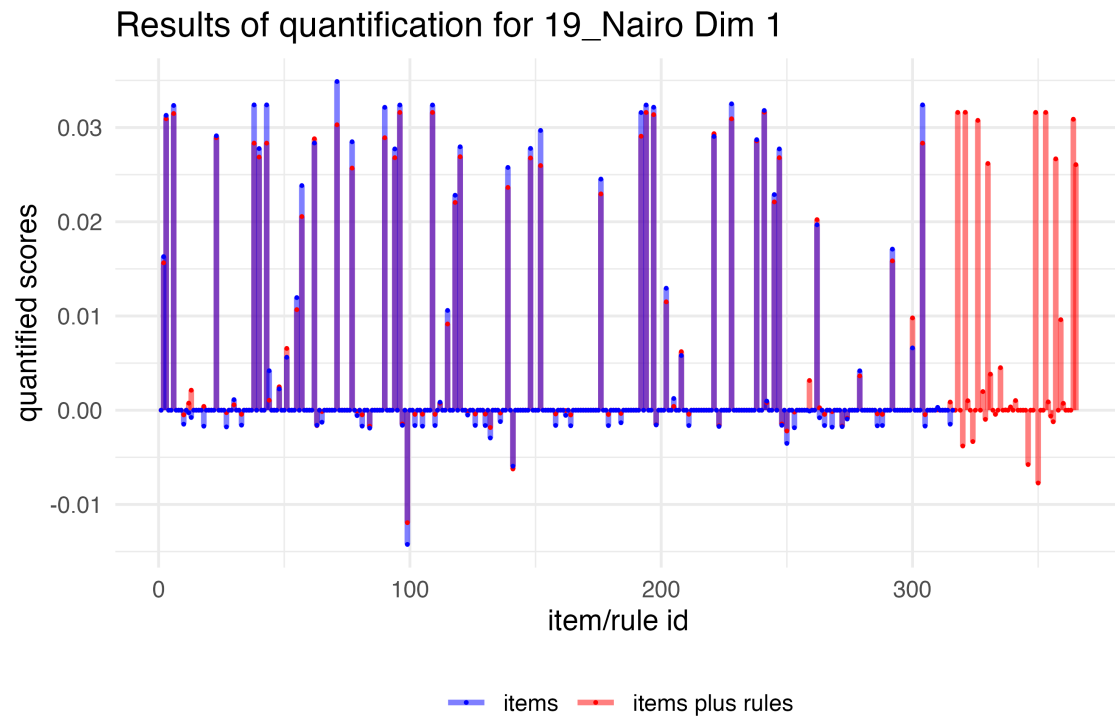


Figure 59. The quantified scores in the first dimension of the Nairo dialect (No. 19 in Figure 1).

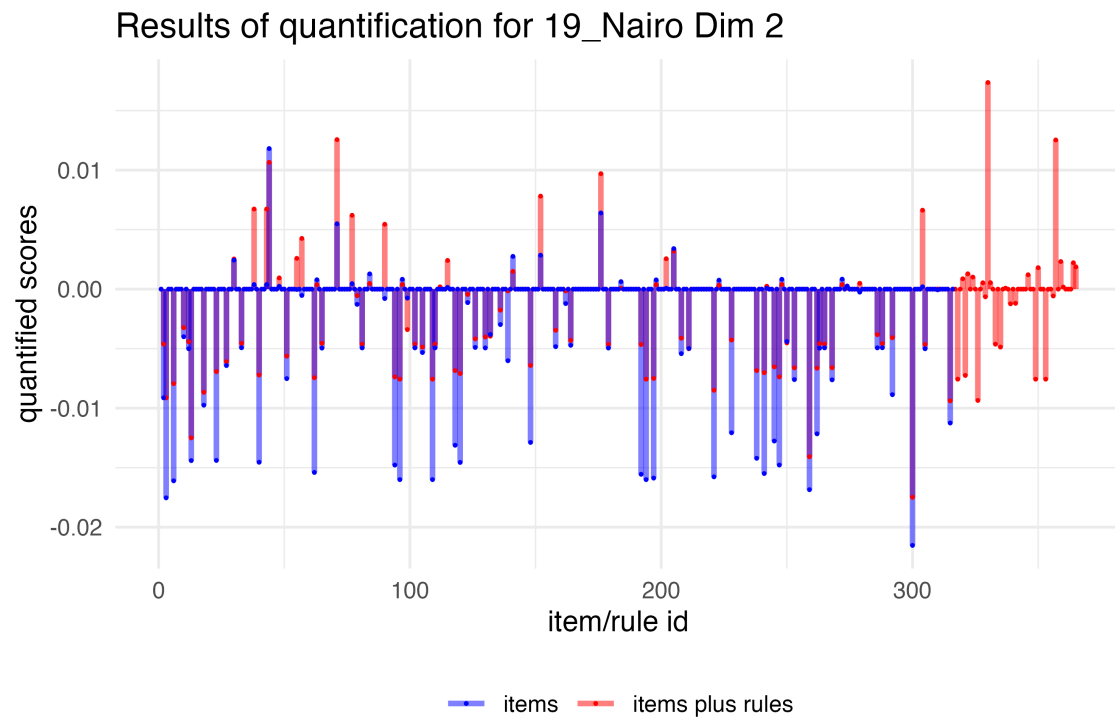


Figure 60. The quantified scores in the second dimension of the Nairo dialect (No. 19).

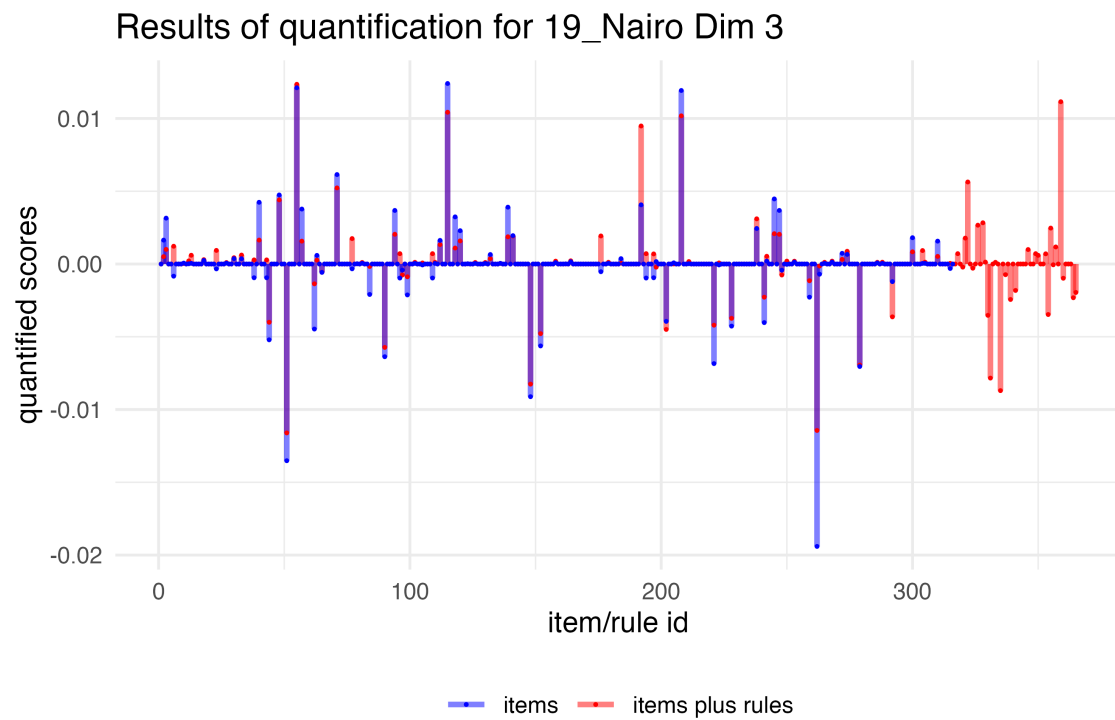


Figure 61. The quantified scores in the third dimension of the Nairo dialect (No. 19).

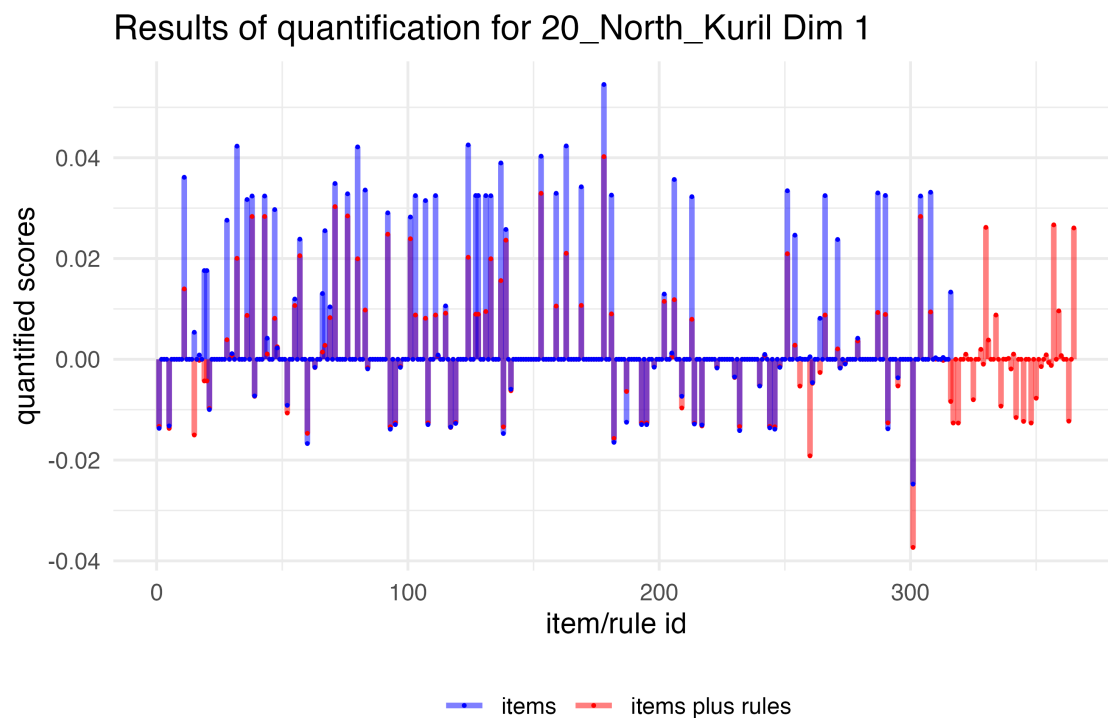


Figure 62. The quantified scores in the first dimension of the North Kuril dialect (No. 20 in Figure 1).

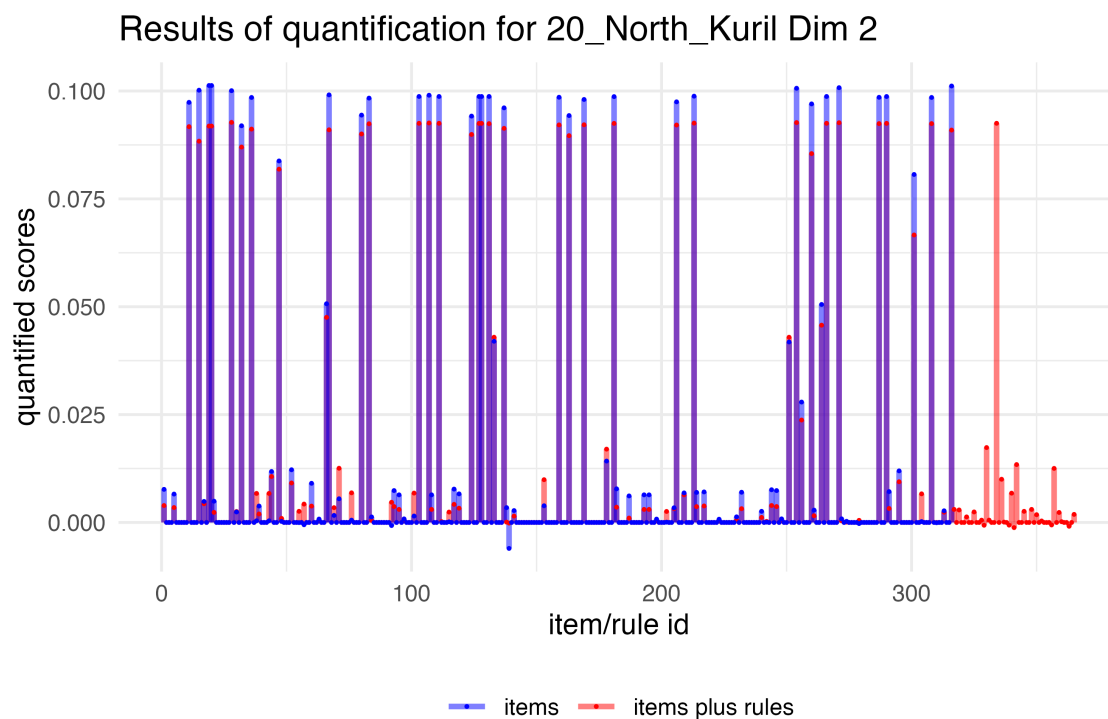


Figure 63. The quantified scores in the second dimension of the North Kuril dialect (No. 20).

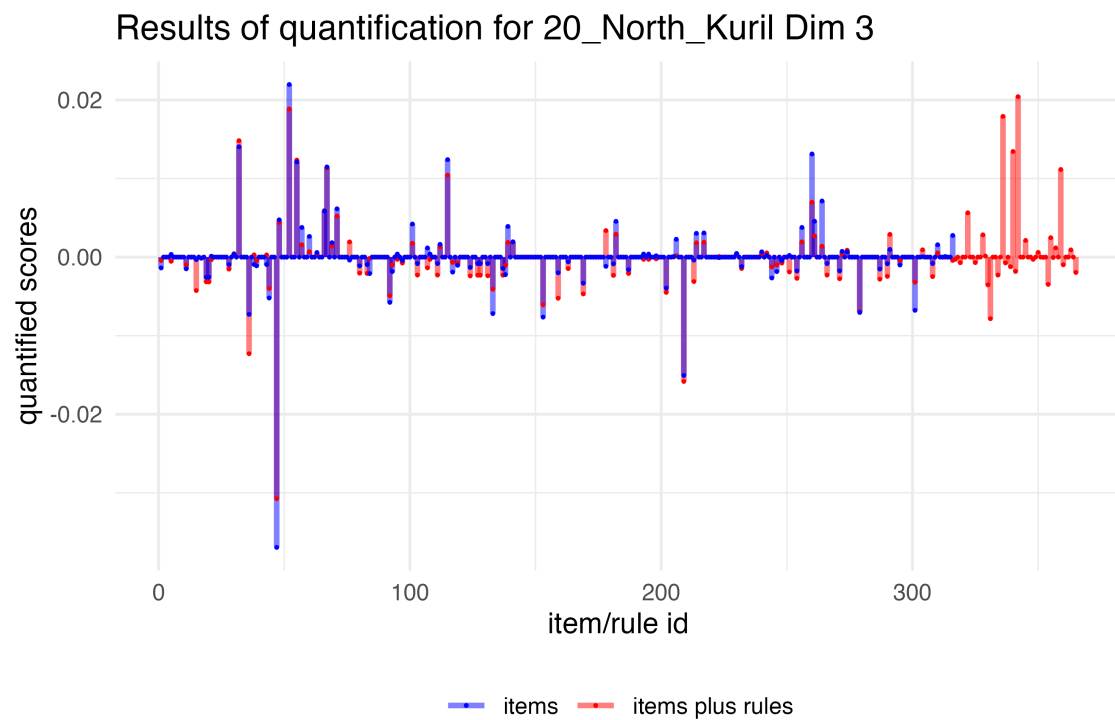


Figure 64. The quantified scores in the third dimension of the North Kuril dialect (No. 20).

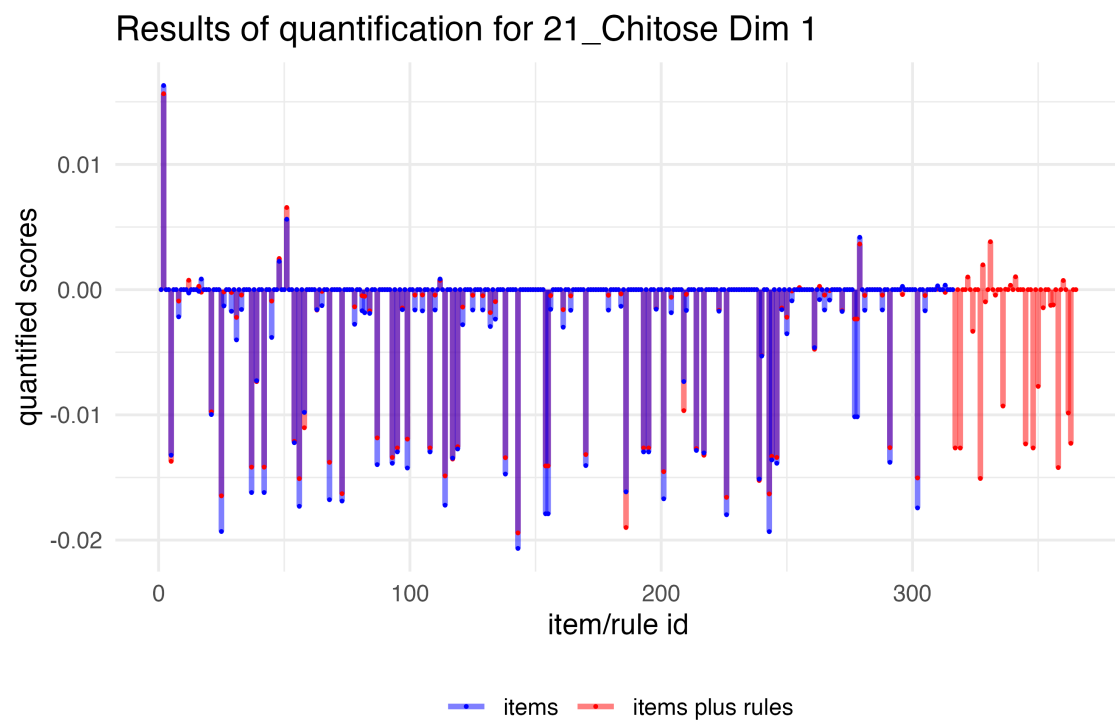


Figure 65. The quantified scores in the first dimension of the Chitose dialect (No. 21 in Figure 1).

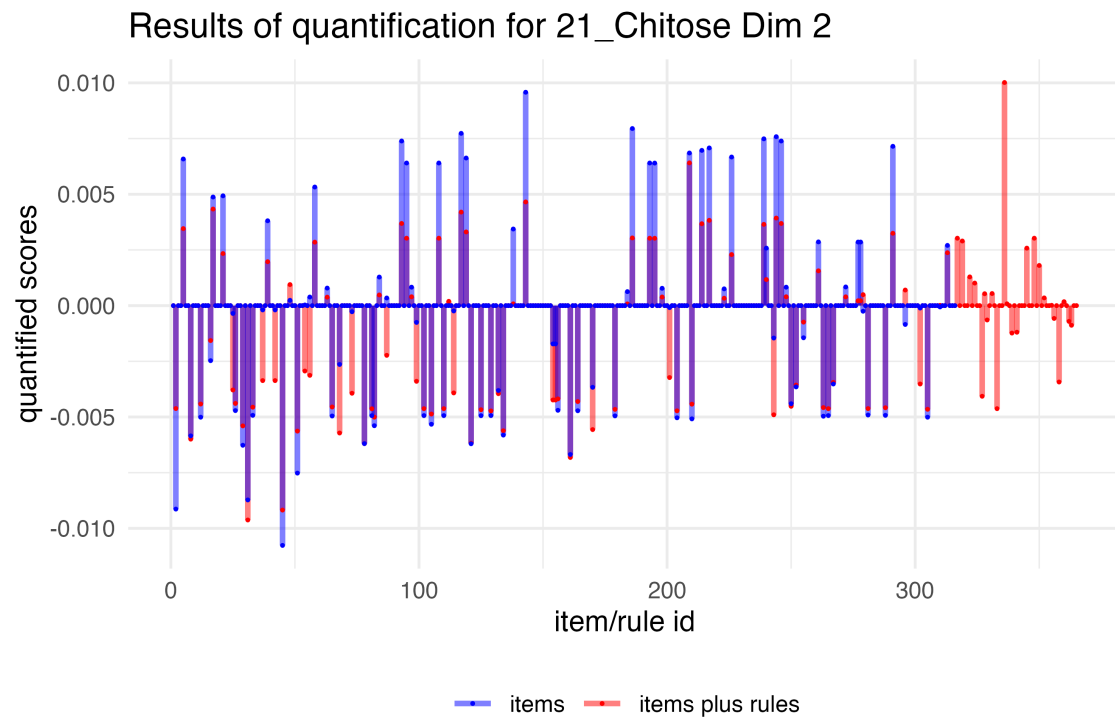


Figure 66. The quantified scores in the second dimension of the Chitose dialect (No. 21).

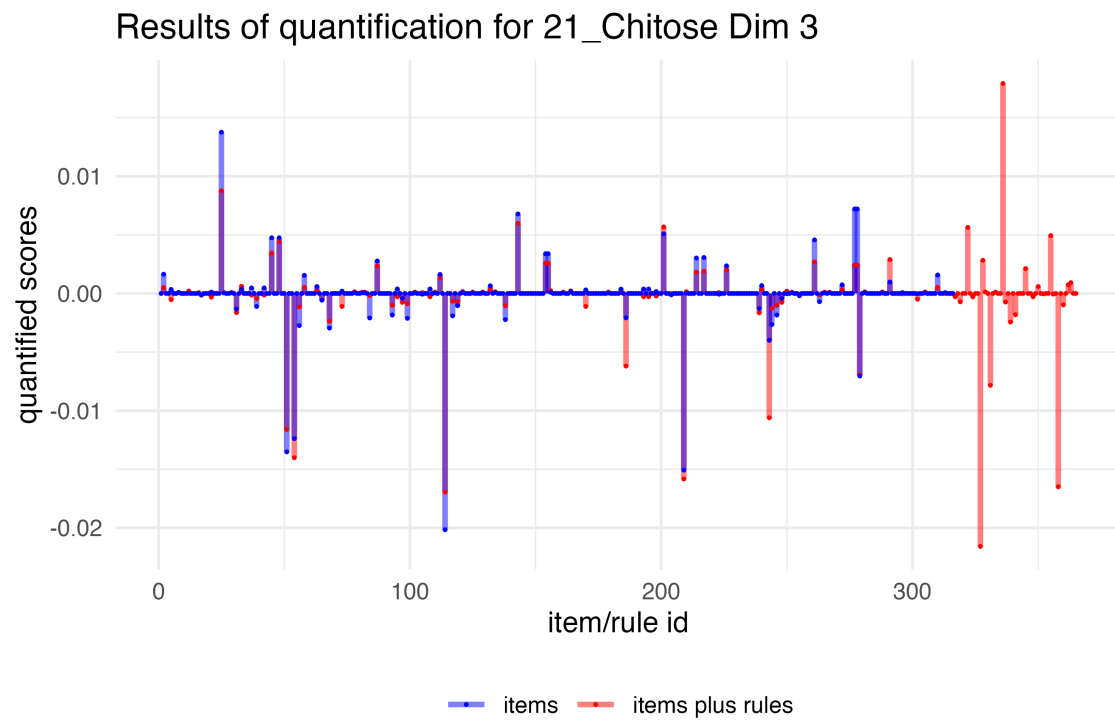


Figure 67. The quantified scores in the third dimension of the Chitose dialect (No. 21).

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.
- 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 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 (in press) Gokei to kisogoi no fukugōteki na deta no kōchiku: ainugo kisogoi no jōhō no saidaika ni mukete [How to build composite data on lexical forms and regularities: extracting maximum information from the basic vocabulary of Ainu dialects]. *Northern Language Studies*, 16.
- 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.
- Hattori, Shirō and Mashihō 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.
- Horst, Paul (1935) Measuring complex attitudes, *Journal of Social Psychology*, 6: 369–374.
- 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>
- Murayama, Shichirō (1971) *Kita Chishima Ainu-go [Ainu language of Northern Kuril Islands]*. Tokyo: Yoshikawa-Kōbun-Kan.
- 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 shikōni 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.
- Richardson, M., and Kuder G.F. (1933) Making a rating scale that measures, *Personnel Journal*, 12: 36–40.
- 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.

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