Chapter 20

Developing the Linguistic Atlas of Japan Database and advancing analysis of geographical distributions of dialects

Yasuo Kumagai
National Institute for Japanese Language and Linguistics, NINJAL

The Linguistic Atlas of Japan (LAJ) was published from 1966 to 1974. A total of 285 items (mainly from the lexical field) and 2400 localities were surveyed. In 1999, we started constructing the Linguistic Atlas of Japan Database (LAJDB) with the aim of preserving the original survey materials (540000 cards) and advancing various uses of the LAJ. To explore the potential of the LAJDB in advancing quantitative approaches to the LAJ, I made some preliminary observations at the locality level concerning, for example, (1) geographical distributions of the frequency of standard forms, multiple answers, and informants' (speakers') comments on standard forms; (2) geographical distributions of degrees of linguistic similarities among localities; and (3) network representation of the degrees of linguistic similarities. These nationwide "high resolution" patterns of 2400 localities showed clear patterns and structures. Comparing these patterns and structures with each other and with extra-linguistic features, such as the network of roads, enables us to examine their relations in detail. The former nationwide, prefecture-level quantitative studies using the LAJ could not achieve such examination. I present some of these preliminary results and discuss several implications for advancing quantitative analysis using the LAJ.

1 Introduction

The *Linguistic Atlas of Japan* (*LAJ*), with six volumes (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966–1974), is the first nationwide Japanese linguistic atlas based on a linguistic geographical survey method. Published from 1966 to 1974, it is one of the basic research materials in Japanese dialectology. Many studies on Japanese dialects have utilized the LAJ—including studies based on an interpretation of each linguistic map, such as Satô (1986), Tokugawa (1993), and Kobayashi (2004),

Yasuo Kumagai. 2016. Developing the Linguistic Atlas of Japan Database and advancing analysis of geographical distributions of dialects. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects*, 333–362. Berlin: Language Science Press. DOI:10.17169/langsci.b81.159

as well as quantitative studies based on an accumulation of maps, such as Takada (1969), Hondô (1980), Kasai (1981), Ichii (1993), and Inoue (2001).

However, quantitative studies that use the LAJ have certain limitations. First, such studies, which used each survey point as a unit of calculation, were restricted to examination of small areas. Next, the nationwide studies were mainly based on prefecture-unit calculations. For example, nationwide geographical distributions of standard word forms in the LAJ have been one of the most analyzed subjects in quantitative studies using the LAJ (e.g., Inoue 2001), but such studies are based on prefecture-level calculations,[1] mainly due to difficulties in generating LAJ data at the locality level (e.g., Hondô 1980: 485, 498). In these studies, researchers prepared their data individually by reading the printed maps. It would have been very laborious to prepare data in such a manner, and it would have been difficult to achieve accuracy everywhere. The lack of digital data certainly restricted the methods of quantitative studies utilizing the LAJ.

We have been developing the Linguistic Atlas of Japan Database (LAJDB), aiming to preserve the original survey materials and advance the utilization of the LAJ. The LAJDB provides data from 2400 survey localities. Calculations at the locality level enable researchers to observe "high resolution" geographical distribution patterns (approximately 50 times the resolution of 47 prefectures). These "high resolution" patterns enable researchers, for example, to trace various diffusion routes that former studies could not detect. Compared with the former studies using the LAJ, studies using the LAJDB can compare such diffusion routes with extra-linguistic features, such as road networks, to examine the relation among them in detail. The LAJDB provides not only the geographical distribution data of word forms but also the original survey card images. This feature proves useful for advancing the utilization of the LAJ. In section 2, I provide an overview of the LAJ and LAJDB. In section 3, I describe some results of preliminary observations to elucidate the potential of the LAJDB.

2 LAJ and the LAJ Database (LAJDB)

The survey for the LAJ was conducted from 1957 to 1965 by the National Language Research Institute (NLRI), which preceded the present National Institute for Japanese Language and Linguistics (NINJAL). A total of 285 questionnaire items (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 105–118), mainly pertaining to

[1] Japan has 47 prefectures.

the lexical field (nouns, verbs and adjectives), and 2400 localities[2] were surveyed by 65 fieldworkers through personal interviews. In principle, one male informant (speaker) born before 1903[3] and born at the given location (or, at least, who spent time there without interruption from the age of 3 to 15) was chosen as informant.[4] As far as possible, those representing the general trend of occupation in their locality were chosen. Here, we may note that approximately 80% of all localities are agricultural communities (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 22, 42). Consequently, for practical reasons, male informants were chosen at 2392 localities (99.7%) and female informants were chosen at eight localities. Concerning informants, 97% were born between 1879 and 1903. Moreover, 63% of informants were engaged in agricultural work, 21% in commerce, and the rest in five other occupation categories (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 24–26, 42, 43). Almost all informants of the LAJ can be described as "NORMs," that is, non-mobile, older, rural males (Chambers & Trudgill 1998: 29). It was reported that "on an average about 6 localities were surveyed in every 1000 square kilometers, or an average of about 12 kilometers separates each surveyed locality" (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 41). The LAJ covers the whole of Japan, and the maximum geographic distance for a locality pair is approximately 2960 km. Figure 1 provides an example of maps from the LAJ.

During the survey period of the LAJ, after completing the questionnaire, each informant's answer was copied to a separate card by fieldworkers. These cards, which number approximately 540000 (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 38, 43) and represent the materials of each answer from each locality, were used as original survey materials while editing the LAJ. On these cards, we can see the original phonetic transcriptions by the fieldworkers, informants' comments, fieldworkers' and editors' notes and so on, which were utilized in editing the

[2] The number of surveyed localities was not equal for each item. The approximate numbers are as follows: Of 285 items, 128 items were surveyed at 2400 localities, 36 items at 2000 localities, 55 items at 1700 localities, 62 items at 1000 localities, and 4 items at 400 localities (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 22, 41).

[3] "One inhabitant was interrogated of each 40,000 people, but since the survey chose only male informants born before 1903, and since we know that there were 4,800,000 males of that age in the whole of Japan (1960 figures), our survey actually reproduces the speech of one out of 2000 of that stratum of the population" (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 41–42).

[4] In principle, those whose residence in the locality had been interrupted by absences longer than 36 months were avoided.

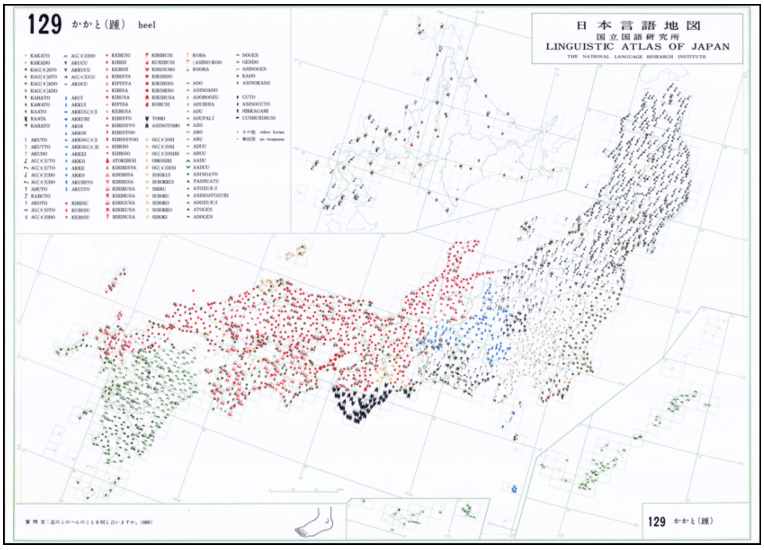

Figure 1: A map from the LAJ (map 129: heel).

maps. However, the published LAJ hardly provided any of this information for users.[5]

In 1999, with the aim of preserving these original survey materials and promoting the utilization of the LAJ, we began constructing the LAJDB (Kumagai 2013a, b).[6] The LAJDB comprises an image database of the original survey cards (Figure 2) and coded data corresponding to the published maps. The items of coded data include (a) locality number (systematically corresponding to degrees of longitude and latitude), (b) item name, (c) linguistic form on the legend of each map, (d) prefecture, (e) number of answers, (f) pattern of multiple answers, and so on.

[5] The editors of the LAJ prepared the materials "Nihon gengo chizu shiryô," which listed the varieties corresponding to every linguistic form presented in the LAJ and the comments of informants and fieldworkers (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 32, 33). These materials were recorded from the original survey cards in handwritten form. They need further proofreading, and the list is partially incomplete. To use this material, it is necessary to check the original survey cards for confirmation. The editors planned to publish the comments and notes (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 44), but this plan was not realized.

[6] After the LAJ was published, the Grammar Atlas of Japan (GAJ, 6 volumes) was published between 1989 and 2006 by the NLRI (NLRI 1989–2006). Surveys were conducted from 1979 to 1982 in 807 localities. In the course of editing the GAJ, the use of computer in publishing the GAJ was developed and the GAJ data was made accessible to the public. However, in the days of the LAJ, computers were in the early stages of development and not available for publishing linguistic atlases.

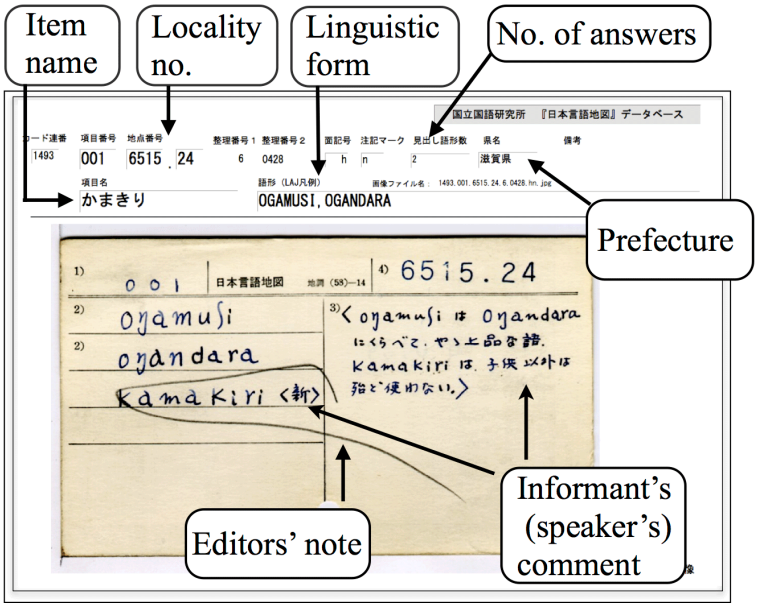

Figure 2: A snapshot of the LAJDB image database.

The original survey card images are linked to each entry of the coded data. While the linguistic forms shown on the maps from the LAJ are the result of the editor's classifications of varieties recorded by the fieldworkers, the LAJ does not provide us with detailed information about the classifications. Combining the coded data from the LAJ and the original survey material card images, the LAJDB allows, for example, tracing of the classifications and interpretations completed by the editors. Moreover, the LAJDB facilitates close examination of the LAJ as a research material, reclassification of linguistic varieties based on other viewpoints, and utilization of informants' comments and field workers' and editors' notes.

Currently, 119 items have been completed, corresponding to 43% of the number of surveyed items (285 items) and 49% of the number of items published as maps (240 items). The progress of the scanning of the original survey cards has reached approximately 90% of the total number of items (Kumagai 2013a: 159).

In addition, we have been preparing the following data for the LAJDB: (1) informant's information provided in the LAJ (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 47–102), such as (a) address (without house number), (b) year of birth, (c) occupation, (d) educational background (number of years), (e) absence from the locality (number of months), (f) experience of military service, (g) sex, (h) name

of interviewer (fieldworker), (i) year of survey, and (j) questionnaire used; and (2) digital maps—in shapefile format, a standard file format for geographic information system (GIS) software—based on the "Introductory maps" from the LAJ compiled by the editors, including (a) basic maps from the LAJ, (b) topographical maps (showing mountain systems, river systems, etc.), (c) main roads in the Meiji period (approximately 1885), (d) the boundaries of the feudal domains during the Edo period (1664), and so on.

3 Some preliminary observations from using the LAJDB

3.1 Dataset for preliminary observations

As previously mentioned (see footnote 2), the number of surveyed localities for the LAJ was not equal for each item. Therefore, to explore the possibilities of the LAJDB, I selected 55 items from the LAJDB, in progress, with 2400 ± 1 survey points. We will call this dataset LAJDB55 data. The item numbers selected for LAJDB55 are as follows: (001), (005), 006, 007, (012), 032, 036, 038, 039, 048, 051, 052, (056), 057, 059, 060, 063, 064, 066, 067, 072, 076, 083, 089, 103, [104], [105], 110, 111, 116, 118, 119, (122), 124, (127), 129, 148, 149, (165), 174, 179, 185, (186), (187), (188), 191, 194, 200, 214, 215, (216), 219, 220, 221, and 223.

Furthermore, I created a subset—LAJDB42—of LAJDB55 to explore the distribution of the standard forms. In most cases, an item has one standard form; however, two or more forms were occasionally recognized as standard by the LAJ's editors. In the above item numbers, those enclosed in parentheses are the ones in which two or more standard forms were recognized by the editors. No standard forms were explicitly stated for the items enclosed in square brackets []. For convenience in processing standard forms, I omitted the 13 items in parentheses from LAJDB55 and acquired them from LAJDB42. Using the LAJDB55 and LAJDB42 datasets, I made some preliminary observations.

3.2 Geographical frequency distributions of standard forms

Leading studies using quantitative analysis of the LAJ have been performed by Fumio Inoue (Inoue 2001, etc.). Inoue (2001), in a collection of his research of approximately 20 years, analyzed the usage rates of the standard forms of 84 LAJ items, summed up by prefecture (47 prefectures and Hachijô island, which belongs to Tokyo metropolitan area but whose dialect differs significantly from that of Tokyo). This data was originally prepared by Hisako Kasai (1981) by hand.

Inoue input the data for quantitative analysis (Inoue 2001: 89). This data is known as the Kasai data. Inoue (2001) analyzed the Kasai data via multivariate analyses, quantitative techniques, and so on, in order to explore dialect areas, the geographical diffusion of dialects, the distribution of standard forms, and so on.

Figure 3[7] shows the distribution of the usage rates of the standard forms of Kasai data on the map. With the Tokyo metropolitan area as the peak, the usage rate gradually declines toward the periphery, resembling a wave-like diffusion with Tokyo at its center. Hokkaido is an exception as it was a new settlement with people from mainland Japan.

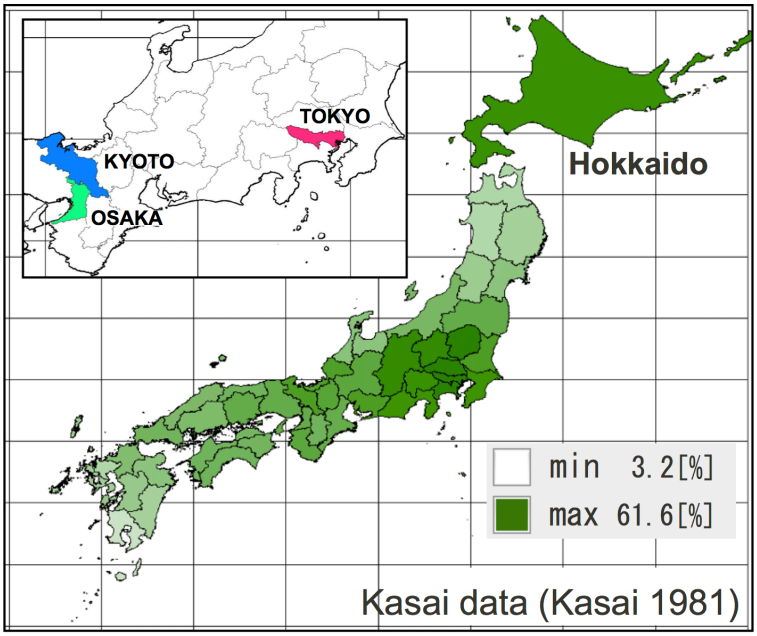

Figure 3: Usage rates of standard forms of 84 LAJ items, summed up by prefecture.

It should be noted that the gridlines of the map are drawn based on the locality number system of the LAJ. The locality number system was based on topographical maps with a scale of 1:50000. Each block in the grid corresponds to 100 topographical maps of a 1:50000 scale. The size of each block in the grid is 2°30' east–west, 1°40' north–south. The gridlines that appear hereafter are similar.

[7] Figure 3 was created based on Kasai's (1981) calculation. Okinawa, located in the southernmost part of Japan, is not displayed in this figure due to space constraints.

Figure 4 shows the usage rates of the standard forms according to LAJDB42 data based on the same calculation method as Kasai data, which shows similar distribution of Kasai data. In Figure 5, the usage rates of the standard forms of Kasai and LAJDB42 data are plotted in the descending order of the values of Kasai data for comparison. Figure 5 indicates that Kasai data and LAJDB42 data show a very similar pattern overall. Figure 6 shows the geographical distributions of the frequency (GDF) of the standard forms of LAJDB42 data by 2400 survey points. The GDF of standard forms was calculated by simply totaling the number of standard forms at each locality in the dataset. Patterns can be observed in detail, which could not be obtained from the prefecture-unit calculations of former studies. These "high resolution" patterns obtained using the LAJDB enable us to observe the diffusion routes more precisely. Comparing these distributions with the road networks, which play an important role in dialect diffusion, reveals interesting relationships.

Figure 7 shows the main roads in Japan in approximately 1885 (Honshu, Shikoku, and Kyushu areas). This map was created based on "Introductory Map V,"[8] a road map of the modern period, in the LAJ. An explanatory note from the LAJ states that this map provides an overview of land transport at the time during which the informants were growing up. This historical map is useful for comparing land transport—an important extra-linguistic factor—with dialect distributions. (Certainly, other means of transportation existed, but they were excluded as items for future incorporation into studies of the LAJ.) The thick purple lines denote the national roads, and the thin blue lines denote the prefectural roads.

Now let us see the relation between the main roads and geographical distributions of the frequency of standard forms. To illustrate this relation more precisely, Figure 8 focuses on the central part of the Honshu area, which includes Tokyo

[8] Here, it must be noted that "Introductory Map V: The main roads in Meiji period (around 1885)" aimed to provide an overview of the relationships between the surveyed localities and road networks. Roads were drawn on the basic map of the LAJ with reference to 1:200000 scale maps compiled via the Army Land Survey conducted by the General Staff of the Imperial Japanese Army (153 maps compiled and published from 1884 to 1893). The editors of the LAJ selected the national roads and prefectural roads that form the maps. The explanatory note stated that although there may be some roads that had been planned but not realized, "Introductory Map V" contained the most important roads around 1885. Figure 7, a digital map, was made by tracing the roads on "Introductory Map V," and the projection system of the map from the LAJ was not explained in its documentation. Thus, the map shown in Figure 7 is an approximation and involves some deviation (This map will be checked against the original compiled maps). Nonetheless, it is valuable and useful for explorative observation. Further, in the following observation, I consulted some related maps, books, and so on to confirm the observations. See also footnote 9.

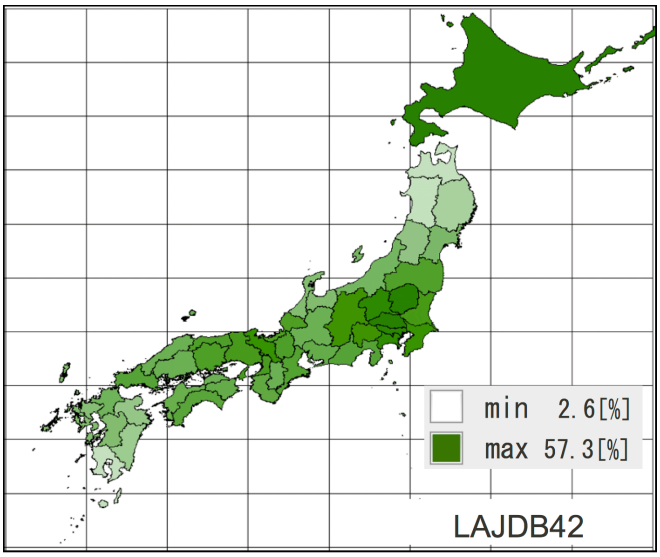

Figure 4: Usage rates of standard forms of LAJDB42 data, summed up by prefecture.

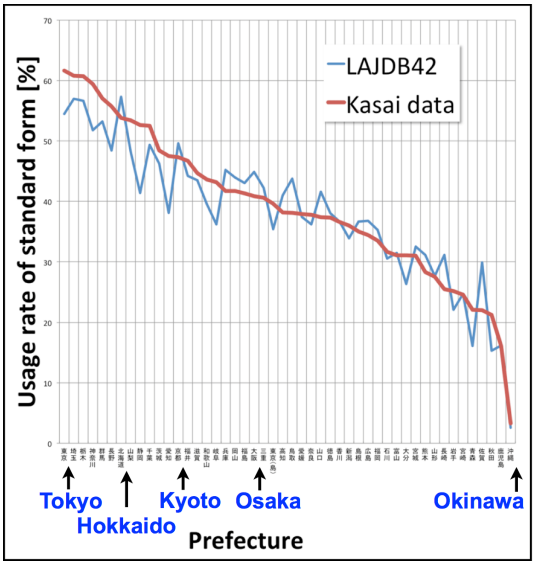

Figure 5: Comparison between Kasai data and LAJDB42 data.

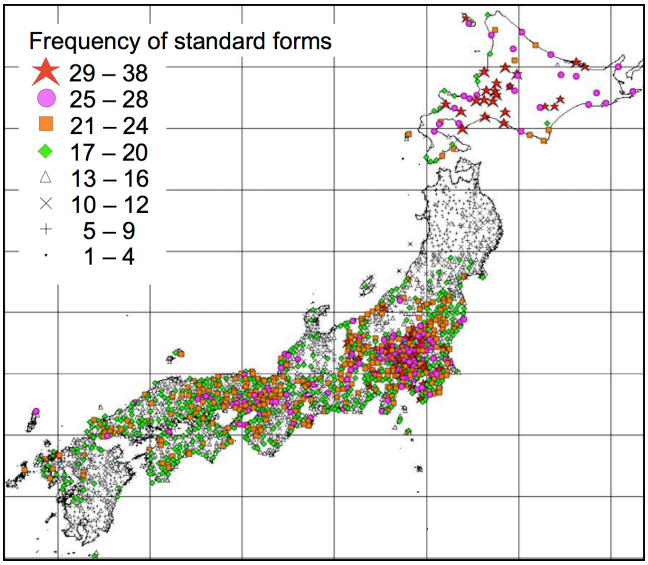

Figure 6: Geographical distributions of frequency (GDF) of standard forms from LAJDB42 data.

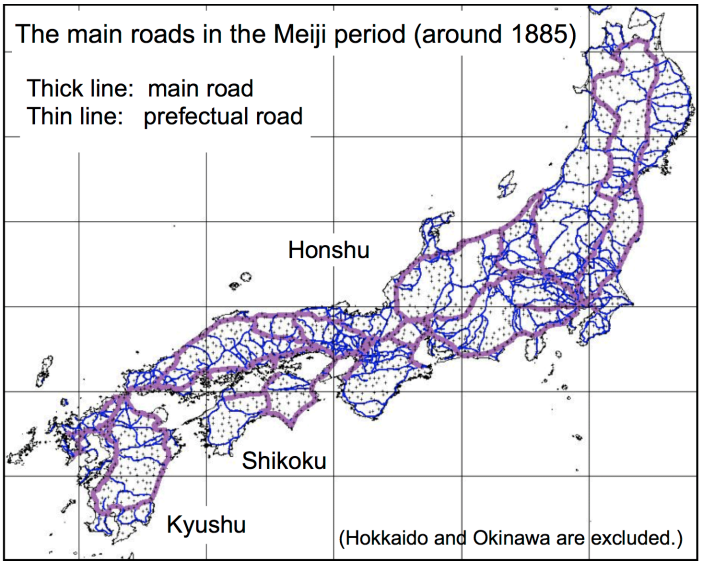

Figure 7: The main roads in the Meiji period (around 1885, Honshu, Shikoku, Kyushu areas).

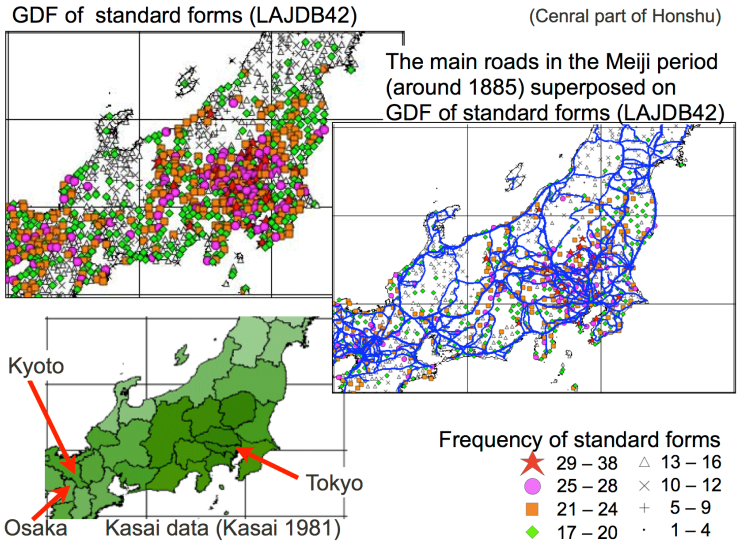

Figure 8: Comparison between Figure 6 and Figure 7 (Central part of Honshu).

(the current capital city) and Kyoto (the former capital city). The roads are superimposed on the distributions of the frequency of standard forms. Interesting observations can be formed regarding the relation between the distributions and the roads.

The map in Figure 9 includes Tokyo (current capital), Kyoto (former capital), and Osaka (large commercial city). The roads connecting Tokyo, Kyoto, and Osaka are very important. There are main roads ("Kaido"), side roads ("Wakikaido"), and others. Tokaido and Nakasendo are the two major main roads connecting these principal cities (Figures 11 and 12). Based on the Kasai data mentioned previously (see Figure 3), Tanaka (1991: 184) observed that the distribution of relative high frequency usage rates along the Tokaido route is interesting and noteworthy. The Nakasendo route runs through the mountainous areas, and Tokaido was the route used by feudal lords in the Edo period (17th century to the middle of the 18th century) to travel to Edo (present Tokyo). The Japanese tend to consider the Tokaido route rather than the Nakasendo route as the major road connecting Tokyo and Kyoto. However, based on LAJDB42 data, the Nakasendo route stands out. Localities with a high frequency of standard forms are plotted along the Nakasendo route. Comparisons drawn between the Tokaido and Nakasendo routes produce interesting results. Future studies on transportation history facilitate deeper insights pertaining to this observation.

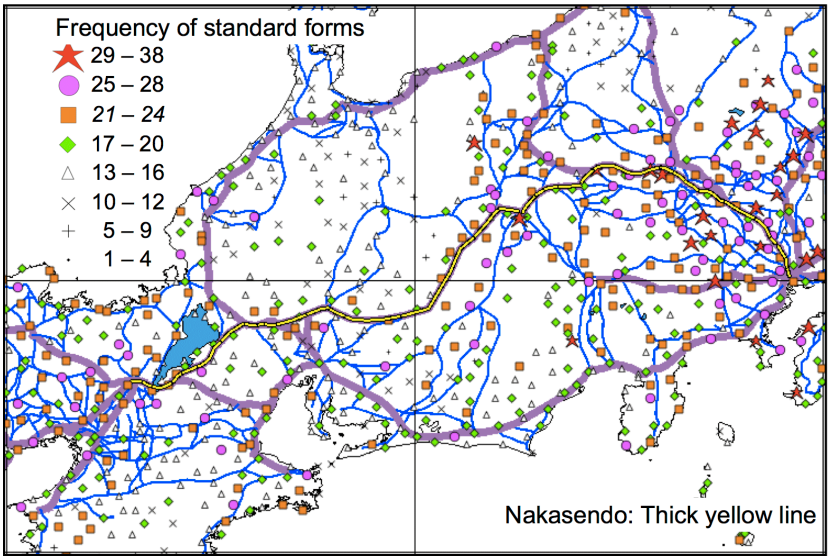

Figure 9: GDF of standard forms superimposed on the main roads in the Meiji period [Nakasendo].

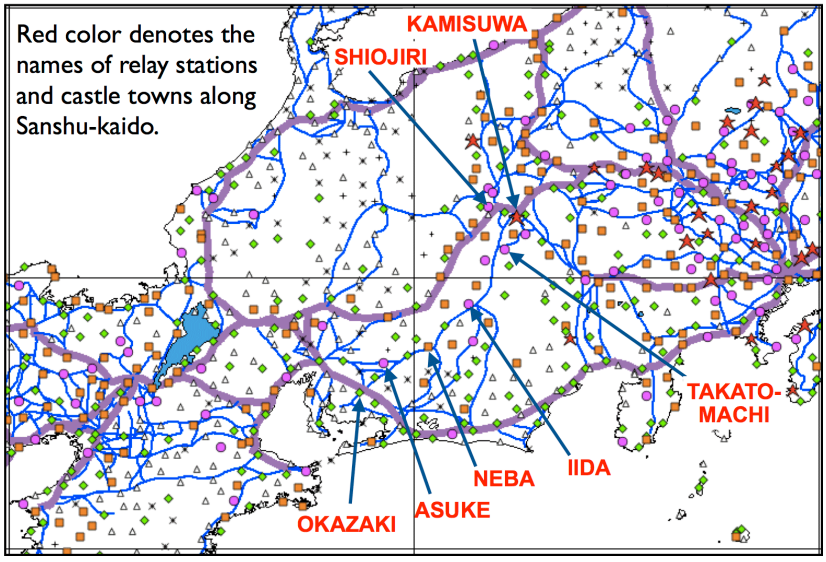

Figure 10: GDF of standard forms superimposed on the main roads in the Meiji period [Sanshu-kaido].

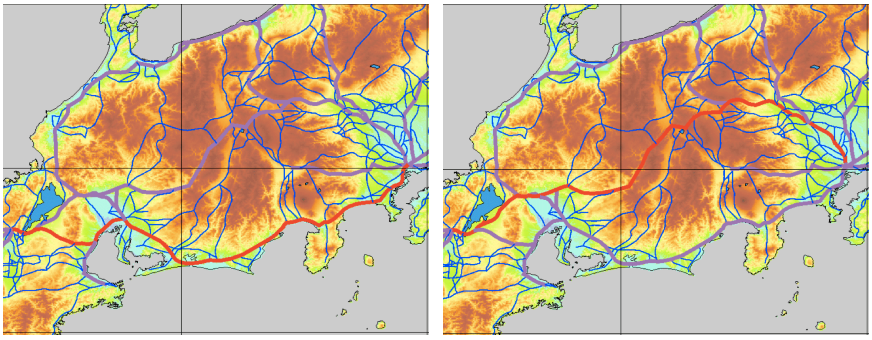

Figure 11: Left: Tokaido route. Right: Nakasendo route.

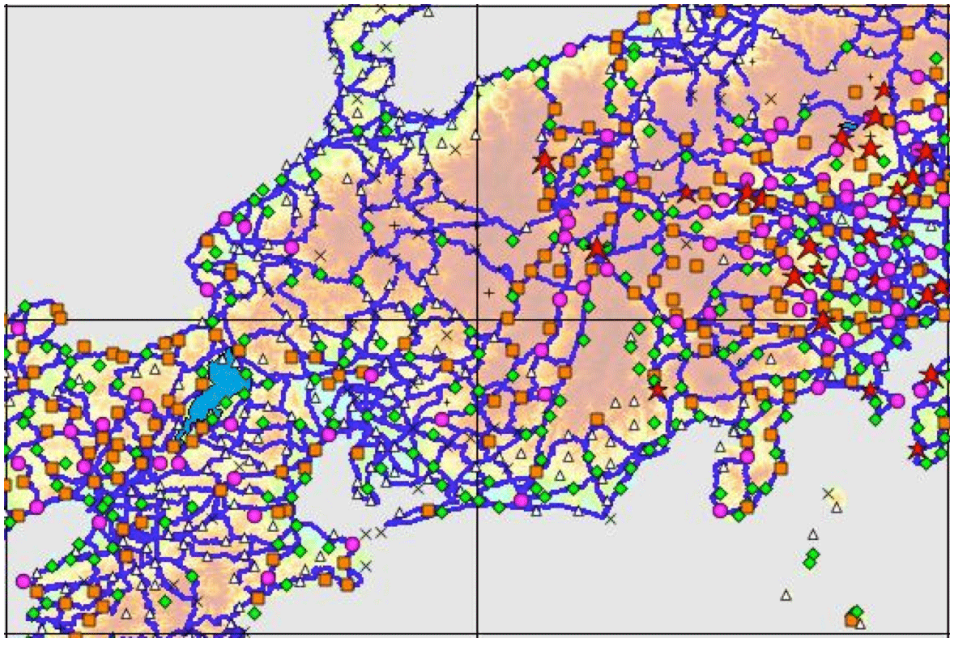

Figure 12: GDF of standard forms superimposed on current road network.

Another interesting example is Sanshu-kaido, a side road of Nakasendo. Sanshu-kaido is a route that connects Shiojiri, Iida, Neba, Asuke, and Okazaki (Figure 10). Similar to the Nakasendo route, Sanshu-kaido appears prominent, with localities with a high frequency of standard forms observed along this route.[9] Sanshu-kaido is not a major road; instead, it was developed as a road for transporting goods. Further systematic observations and analysis should lead to more interesting findings. Notably, these observations were not possible based on the former prefecture-unit calculations of the LAJ.

3.3 Geographical frequency distributions for multiple answers

In some localities, two or more linguistic forms were recorded. These multiple answers play an important role in the interpretation of maps, as they form relations between language contacts, diffusions, and changes. Inagaki (1980) provided some observations about multiple answers on a few maps from the LAJ, and Inoue (2004) noted the importance of these multiple answers and examined their position in the process of diffusion of standard forms.

Few quantitative studies examine the distribution of multiple answers, and the actual status in the LAJ was only partially examined. However, such studies can be easily conducted using computerized data. Figure 13 shows the geographical distributions of the frequency of multiple answers. LAJDB55 data is used here. This distribution contains all items including standard forms. It shows a significant distribution and is not distributed randomly all over Japan. Figure 14 shows the localities color-coded according to the fieldworkers. As Fumio Inoue noted,[10] it is probable that some fieldworkers tended to record more multiple answers, while others tended to record fewer. As a rule, the LAJ survey was designed to maintain uniformity[11] among fieldworkers; however, it is important to

[9] Figure 12 shows the geographical distributions of frequency of standard forms superimposed on the primary route at present (around 2010). This map is prepared for double-checking. The localities with a high frequency of standard forms along Sanshu-kaido coincide with the route better in this map (see footnote 8). It must be noted that new roads are sometimes built along old roads and other times are not. As a whole, this observation also supports the observation above. Road network data: Geospatial Information Authority of Japan (2011); Global Map Japan in Global Map ver. 2.0. Elevation data: Geospatial Information Authority of Japan (2000); Global Map Japan in Global Map ver. 1.0.

[10] A comment by Fumio Inoue, recorded in Inagaki (1980: 6).

[11] During the LAJ survey, to maintain uniformity in the fieldworker's surveys, various attempts were incorporated into the survey design. For example, "to assure a greater uniformity in the questioning, one of the members of the directing dialect bureau from Tôkyô, accompanied the local fieldworkers during the survey of one or more of the assigned localities. The technique

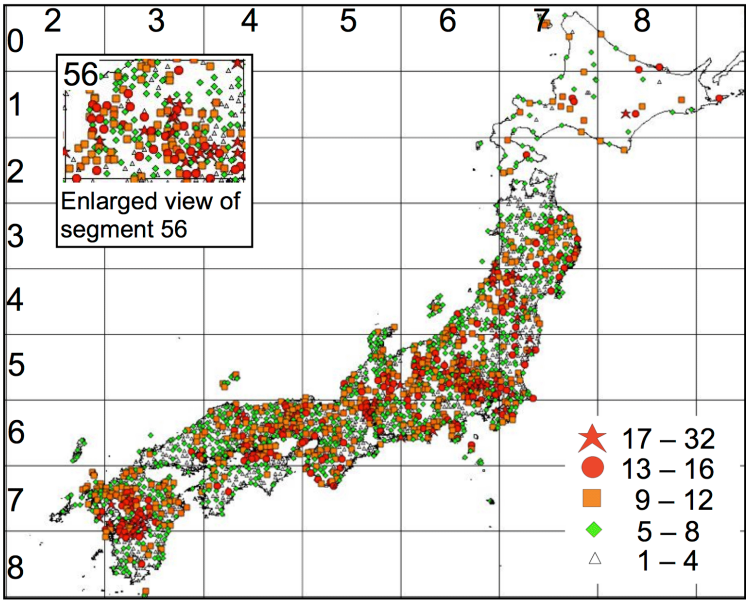

Figure 13: GDF of multiple answers, LAJDB55.

be careful about such risks and verify the observations from multiple perspectives. Accordingly, I compared the distributions of the localities assigned to each fieldworker and the distributions of the frequency of multiple answers. We can see the continuous distribution patterns of the GDF of multiple answers, which spread beyond the boundaries of the fieldworkers' distributions. In other words, we can observe that the boundaries do not limit the continuity of the distribution patterns of the GDF (see, e.g., the enlarged views in Figure 13 and Figure 14).

3.4 The frequency of informant's comments on standard forms among multiple answers

For the LAJ, the editors maintained a principle called the "principle of processing multiple answers." When two linguistic forms were recorded in one locality, both were marked on the map. However, when one of the two forms was the standard language form and, in addition, this fact was noted by the informant—such as in

of selecting an informant and the method of questioning was then demonstrated." Furthermore, "221 localities were surveyed by one of the directors" and "these localities are equally distributed over the whole territory," (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 23, 40–41).

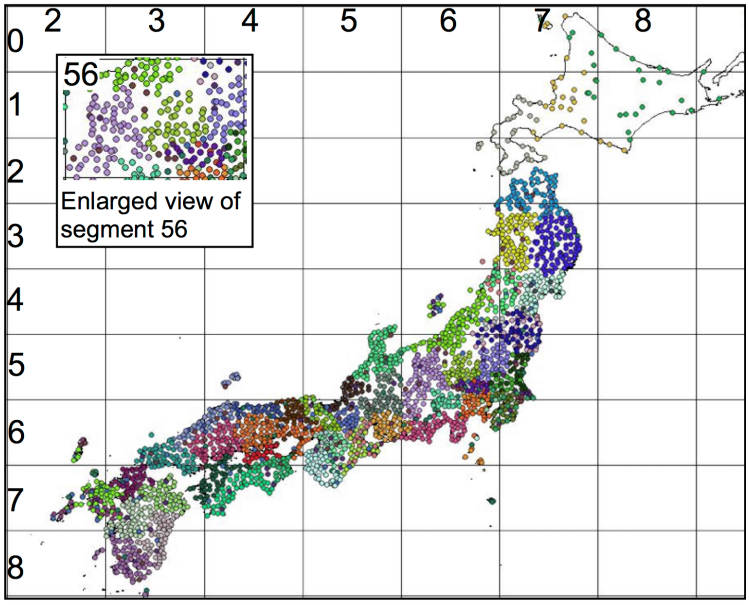

Figure 14: Localities color-coded according to fieldworkers.

the answers "This is a new polite form." or "This is the word used in school."—then the editors would omit the standard forms from a map. This method was followed because the LAJ survey aimed to record informants' personal speech used in their familiar and daily surroundings. Certainly, dialectical forms identical with standard forms were not omitted if there were no informant comments. The principles for processing multiple answers to the LAJ are as follows.

> Further elements of interpretation are given by the informant's comment ("old word," "new form," etc.) or by the fieldworker's notes. These have been helpful for the map interpretation, and they will be published in a later volume.[12] When two linguistic forms have been recorded in the same locality, they have been both marked on the map. When, however, one of the two is the standard language form, and when, in addition, this fact has been noted by the informant ("this is the new polite form," "this is the word used in the school," etc.), in this case only, we have omitted the forms marked this way from the maps (Kokuritsu Kokugo Kenkyûjo (NLRI) 1966: 44).

[12] The publication of comments and notes was not realized. Also see footnote 5.

To study the multiple answers, the omitted standard forms that informants had commented on (e.g., "This is new," and so on) are important. In the course of compiling the LAJ, the editors assumed that the standard forms that were commented on as "new" were distributed randomly. However, using one LAJ item as an example, Satô (1986: 152–153) plotted the omitted answers on a map and found that by adding the omitted answers, the distribution pattern of the standard forms became clearer. However, this observation was based on only one item. The real state of the multiple answers of the LAJ is yet to be explored. Fumio Inoue[13] stated that the editors of the LAJ were aware of some regional differences of the standard forms that were commented on as new and considered these differences as interesting. However, the editors did not plot these words and were unsure about their significance. Using the LAJDB, it is possible to analyze the distribution of the word forms omitted from the atlas.

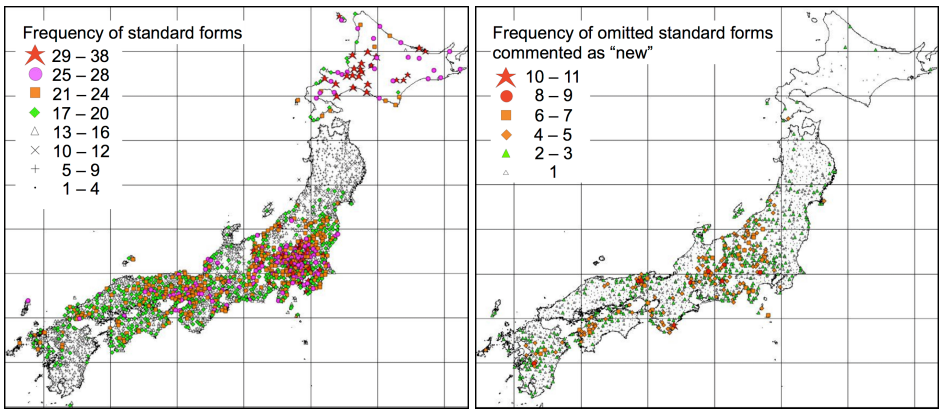

Figure 15: Comparison between GDF of standard forms and omitted standard forms. Left: GDF of standard forms. Right: GDF of omitted standard forms commented as "new".

On the card images provided by the LAJDB, we can see the omitted words and the editors' markings, which signify the application of the principle of processing the multiple answers. In addition, there are lists of localities which record the notes, extracted from the original material cards, with some information about the word omissions performed by the editors. Based on these notes, I formulated data on the omitted word forms using LAJDB42. Figure 15 shows the geographical distribution of the frequency of standard forms and that of omitted standard

[13] A comment by Fumio Inoue, recorded in Inagaki (1980: 5).

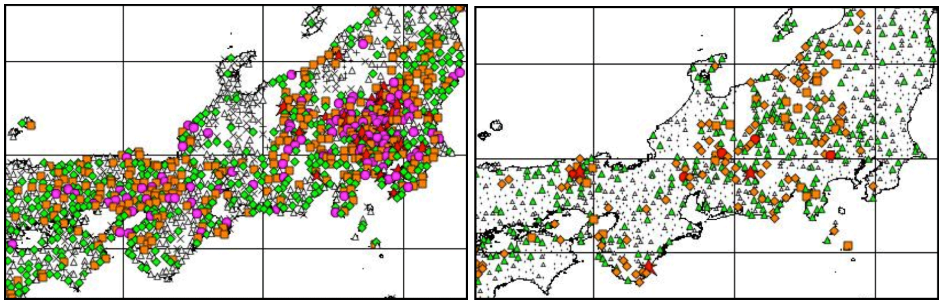

Figure 16: Comparison between GDF of standard forms and omitted standard forms (Central Honshu). Left: GDF of standard forms. Right: GDF of omitted standard forms commented as “new”.

forms for comparison. Is there any relation between the distribution of the standard forms and the distribution of omitted standard forms? To observe this relation more precisely, Figure 16 provides a zoomed-in image. By displaying these two maps alternately as an animation, we compared these two maps visually. For our observation, we focused on the Kinki area and the area surrounding it (Figure 17).

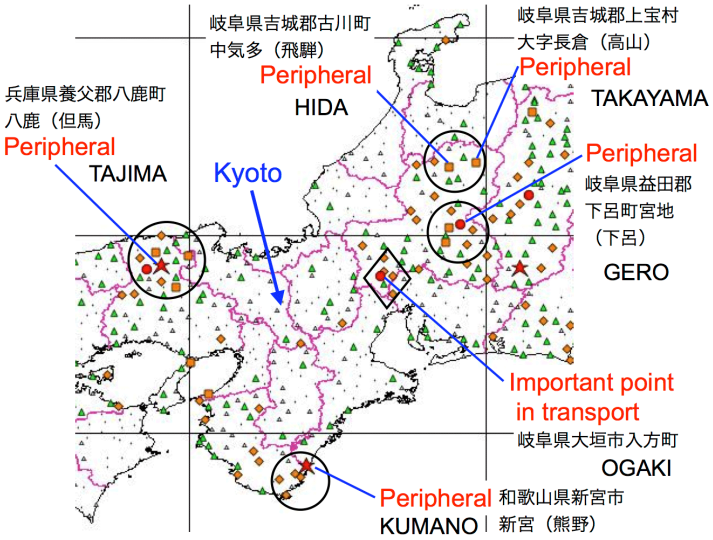

Figure 17: GDF of omitted standard forms commented as “new” (Kinki area).

Figure 17 shows that in areas marked by circles, the areas where standard forms are frequently omitted are surrounded by or adjacent to those where standard forms are very frequent. In areas marked by a diamond shape, the high frequency areas of the two distributions are overlapped. Localities indicated by arrows show the highest frequency in each area marked above. The places marked by circles are the typical peripheral areas. The places marked by diamonds are important areas for transport. In this case, the marked places are clearly separated into two types. Although further investigation is required for other places, these indications are interesting. Possibly, the circled places are at the forefront of the diffusion of the standard forms. More studies should be conducted on areas indicated by diamond shapes.

Figure 18 shows the geographical distribution of the frequency of omitted standard forms superimposed on the main roads in the Meiji period (approximately 1885). Further systematic observations and analysis should present interesting findings. Studies on the history of transportation and other types of knowledge of the regions will be helpful for further studies.

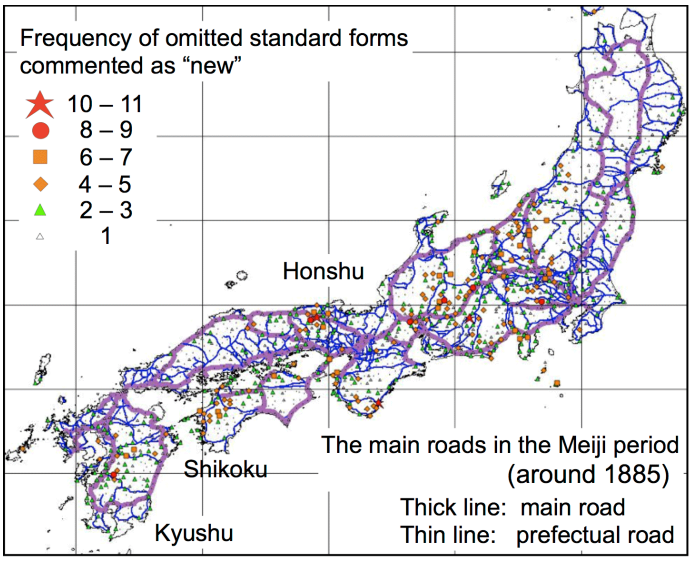

Figure 18: GDF of omitted standard forms superimposed on the main roads in the Meiji period.

3.5 Geographical distributions of degrees of similarity

In this section, the linguistic similarities among the localities based on the LAJDB data will be provided. This information will help present an overall image of the linguistic similarities spread over Japan based on the LAJ. LAJDB55 is used in this section. The previous observations were made mainly with reference to the standard forms. The following maps are based on all word forms including the standard forms. Here, linguistic similarity between two localities is measured by the number of linguistic features shared by the localities. The measure of linguistic similarity used is referred to as NC, that is, a number of common linguistic features. In this paper, NC denotes a number of common word forms (the NC between any two localities is calculated by adding the total number of the same word forms of each item in a dataset). Figure 19 provides some examples[14] of

[14] At the Methods XV conference, the geographical distributions of similarities were represented using animation (total number of frames or maps was 2400). In this case, similarity maps are played after they are sorted by locality number. This animation shows the maps in quick succession. This is an impressionistic form of representation; nevertheless, it allows for observation of reoccurring patterns, transition of patterns, and so on. Such a method of observation should be utilized as an exploratory tool. Figure 19 provides some samples of similarity maps.

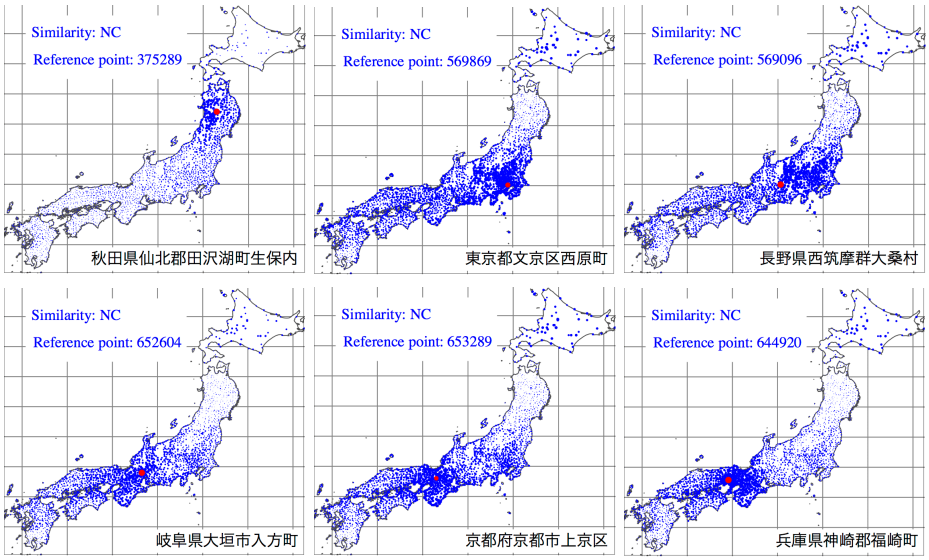

Figure 19: Some example maps of geographical distributions of the degrees of similarity.

geographical distributions of the degrees of similarities. A higher NC value corresponds to a larger radius of circle points. The red points are reference points.

Figure 20 presents similarity maps along the Nakasendo route. A total of 87 points along the Nakasendo route are selected by generating a buffer. Thus, we generated a series of similarity maps of points along a route (Nakasendo). The red line represents the Nakasendo route and the yellow points represent the selected localities. Playing the maps successively from Tokyo to Kyoto as animations

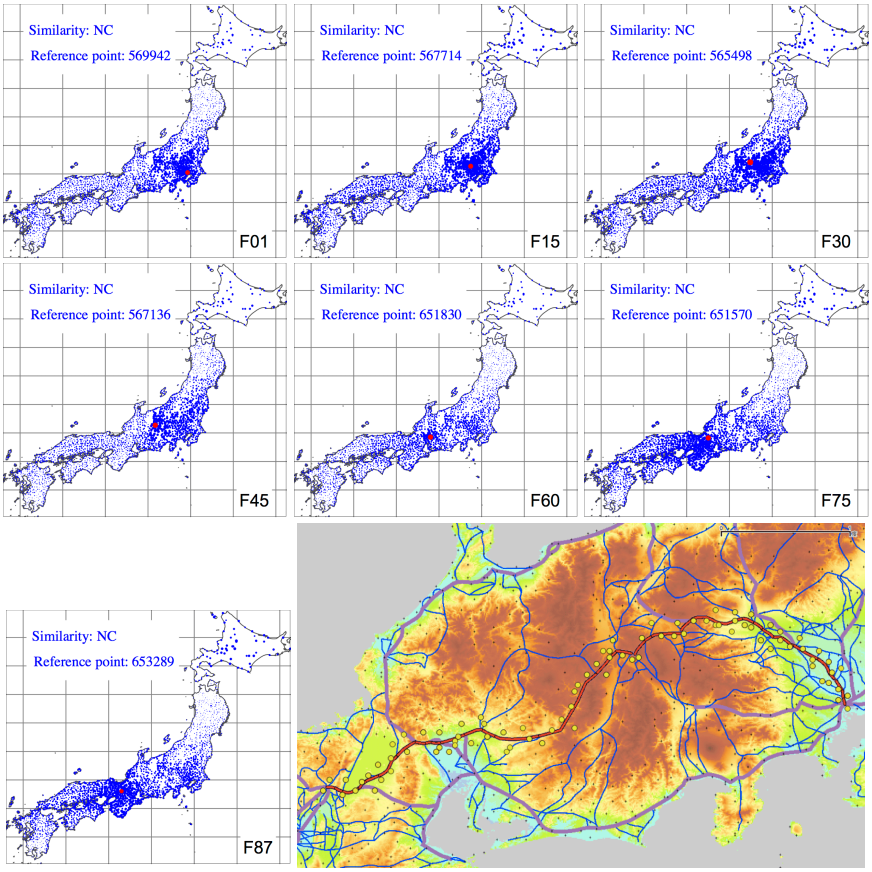

Figure 20: Some examples of similarity maps (frames of the animation) along the Nakasendo route. The number following F is a frame number. F01: Tokyo, F87: Kyoto. Bottom right: The 87 localities selected by buffer along the Nakasendo route (red line: Nakasendo; yellow points: selected localities).

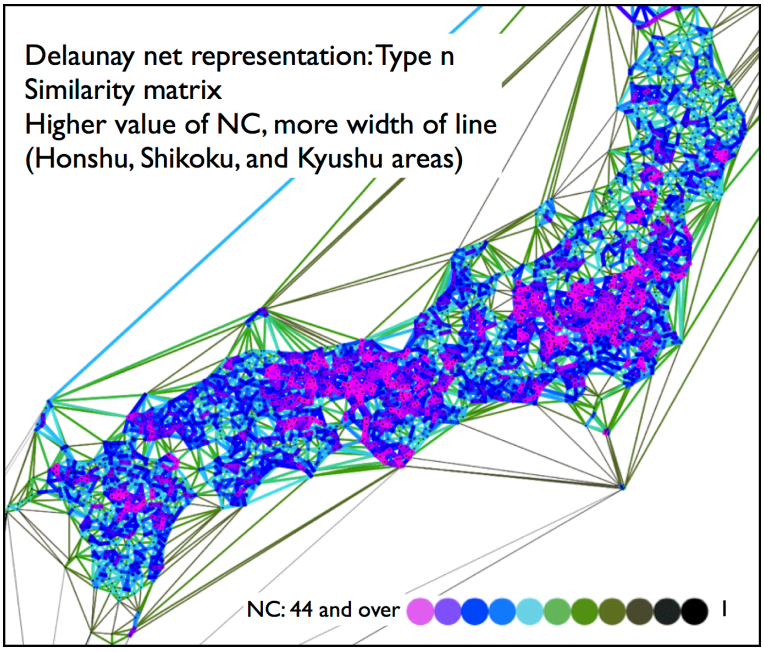

Figure 21: Network representation of degrees of similarity on Delaunay net: Type n.

facilitates observation of the changing patterns along the route. In Figure 20, some examples from the similarity maps (i.e. frames) are shown.

Figure 21 displays an example of another kind of representation of linguistic similarity measured by NC among the localities. Delaunay triangulation — a computational geometrical method to generate a triangular network that connects adjacent points from randomly distributed points on a plane—is used as an approximation to represent continuity among survey points on the geographical space in a formal manner. A network of the points made by Delaunay triangulation is termed as a Delaunay net. We assign a value of NC (number of common linguistic features) to a line which connects two adjacent points of the Delaunay net to visualize the varying degrees of linguistic similarity among survey points distributed on a map. This representation is a network representation on a Delaunay net and is termed type n. Here, "n" represents NC (Kumagai 2013b: 2, 4). Figure 22 presents another example; it provides a network representation of the degrees of similarities on a Delaunay net: type d ("d" denotes distance). The degree of linguistic similarity between adjacent localities is measured by the

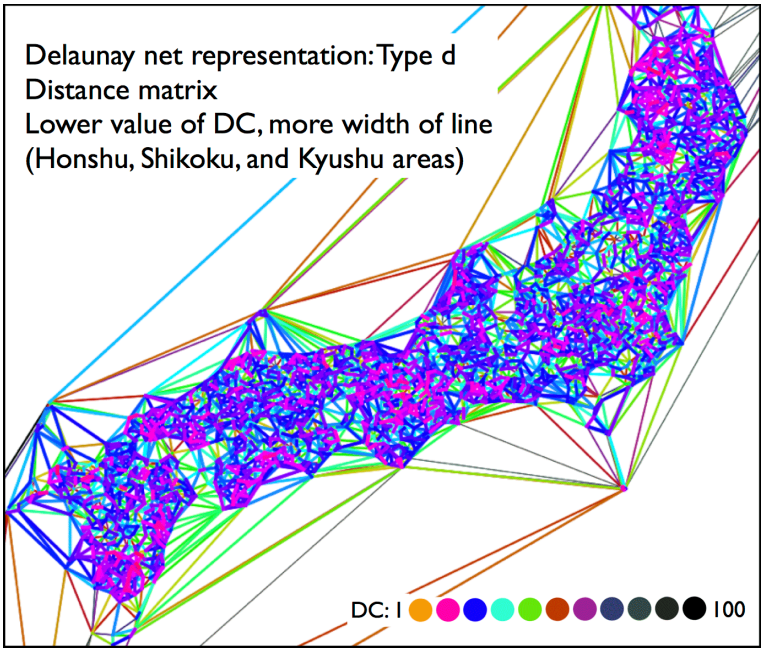

Figure 22: Network representation of degrees of similarity on Delaunay net: Type d.

degree of similarity between the two NC distribution patterns of the localities. A distance matrix is calculated for this purpose. The degree of linguistic similarity between any two localities is measured by the Euclidean distance between the two NC distribution patterns (Kumagai 2013b: 2, 4), and is termed DC. The values of DC are categorized in the range of 100 in Figure 22. Due to space constraints, DC and these maps cannot be discussed in detail; however, clear patterns can be observed on these maps.

Figure 23 shows the NT-1(r)n representation.[15] NT-1(r)n is one of the series of methods we have developed to observe linguistic similarities among localities on a map (Kumagai 2013b: 2). In NT-1(r)n, any two localities that show similarities more than the threshold condition (Lcond) are connected by a line. The red points denote the localities. The measure of similarity used is NC. Any two localities that satisfy the threshold condition (NC $\geqq$ Lcond) are connected by a line. In Figure 23,

[15] In the Methods XV conference, this figure was represented as an animation with changing Lcond stepwise from 48 to 30. In Figure 23, some frames selected from the animation of NT-1(r)n are shown (LAJDB55).

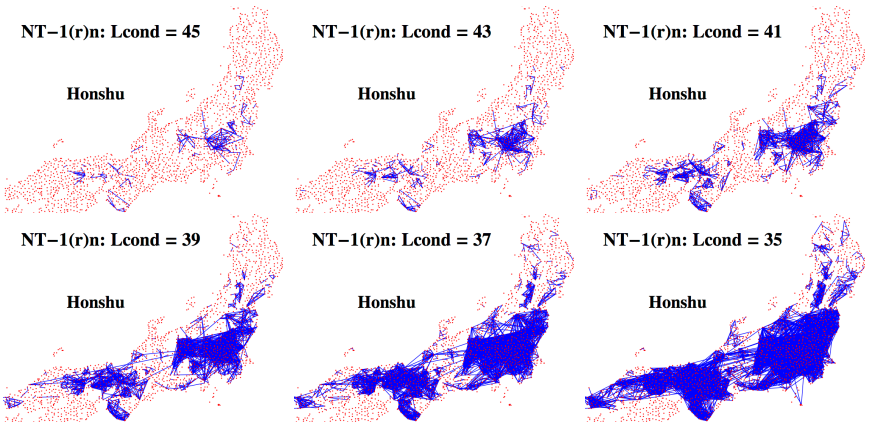

Figure 23: Network representation of NT-1(r)n (Lcond = 45, 43, 41, 39, 37, 35).

only the lines that connect localities inside the Honshu area are displayed to allow focus on observations inside this area. On changing the Lcond from 45 to 30, we can observe how the similar localities are distributed and how the clusters of similar localities grow. Figure 24 displays a superimposition[16] of the network representation NT-1(r)n on the Delaunay net type 2[17] representation (Kumagai 2013b: 6–7). All figures in this section exhibited clear patterns and structures.

It will be interesting to compare these patterns and structures with one another and with extra-linguistic features, such as road networks, to examine the relations among them. The previous observations made on the standard forms must be studied in relation to these observations.

4 Conclusion

We have been developing the LAJDB to preserve the original survey materials and advance the utilization of the LAJ. With 2400 localities, LAJDB data facilitates detailed observations of nationwide distributions, which are not possible

[16] By overlaying two types of representation, we can simultaneously observe the distribution of similarities along the continuity and on the entire map (which is not restricted to neighbors). In transitional zones and homogeneous zones, Nt-1(r) shows the different network structures. By overlaying the two kinds of representation, we can distinguish two types of distribution patterns of similarities, which cannot be distinguished by the representation of the Delaunay net (Kumagai 2013b: 7).

[17] In the type 2 representation of the Delaunay net, a higher NC value corresponds with a lesser line width.

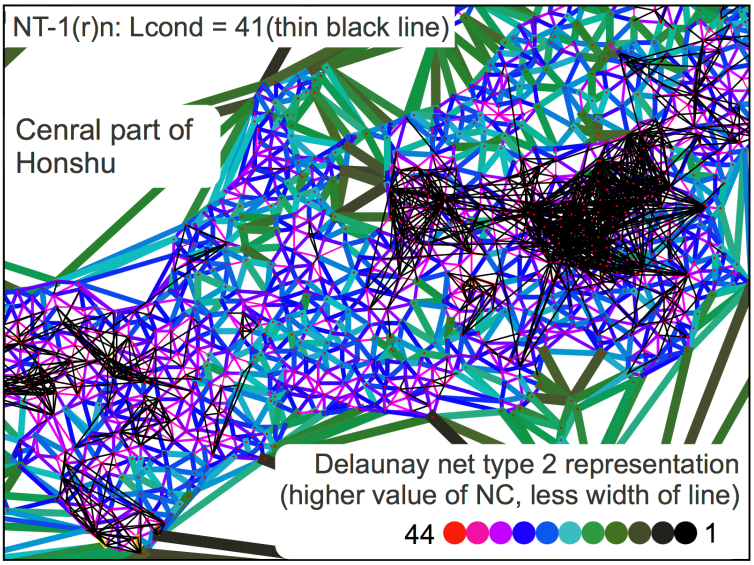

Figure 24: Network representation NT-1(r)n on Delaunay net of type 2 representation.

with the prefecture-unit calculations in the LAJ. In linguistic maps, geographical distribution patterns of each word are usually recognized as distribution areas, that is, planar regions. However, the distributions recognized as planar regions are formed through contact between localities, such as transportation and intercommunication, which refers to contact between individuals (i.e., speakers). By accumulating items for 2400 localities, we will be able to observe the networks responsible for the formation of regions and the phenomena occurring in such networks. Further, we are developing methods for analyzing the geographical distribution data aimed at extracting latent information and finding hidden structure in a manner appropriate to the nature of the data, which will facilitate visualization of the dynamics, flows, and trends of dialectal distribution and understanding of the distribution pattern of dialects in relation to the dynamics of language change (Sibata & Kumagai 1993; Kumagai 2013c,b, etc.). The researchers who conducted the LAJ survey designed many features and devices; however, these tools have not been sufficiently utilized. This shortcoming might partly result from the lack of computerized data, computers, and many other tools which are available today.

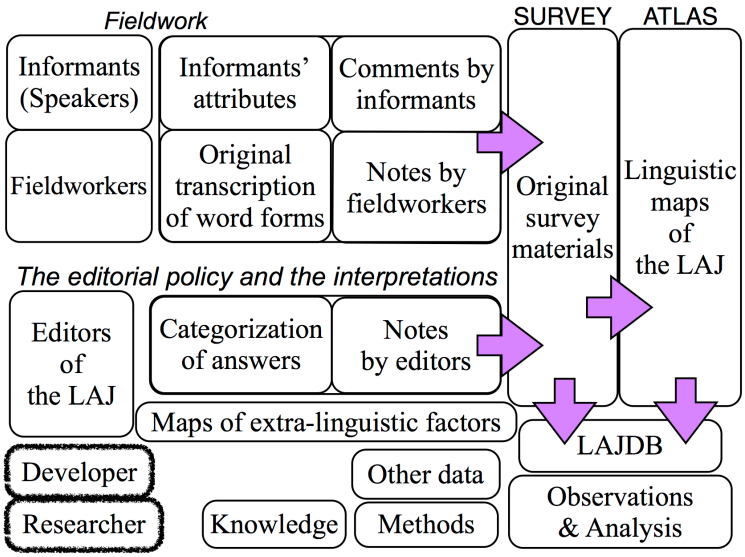

Figure 25: Factors related to the development and utilization of the LAJDB.

Figure 25 illustrates the factors that play an important role in the development and utilization of the LAJDB. All factors relate to developing the LAJDB and advancing the analysis of the geographical distribution of dialects. The digitized data of the LAJ and related information as well as the new methods and perspectives will contribute to advancing the analysis of the geographical distribution of dialects. The LAJDB[18] is expected to be a good tool for utilizing the LAJ and to contribute to advancing the study of the geographical distribution of dialects.

Abbreviations

LAJ	Linguistic Atlas of Japan
LAJDB	Linguistic Atlas of Japan Database
LAJDB42	a subset of LAJDB55
LAJDB55	a subset of LAJDB
NLRI	National Language Research Institute
NINJAL	National Institute for Japanese Language and Linguistics

[18] The LAJDB website is under development (http://www.lajdb.org). This website is currently a work-in-progress, and the LAJDB is only partially open. The site will be updated in accordance with LAJDB progress.

NORM	non-mobile, older, rural males
GDF	geographical distributions of frequency
NC	number of common linguistic features (number of common word forms)
GIS	geographic information system
NT-1(r)n	a method of network representation of linguistic similarities
Lcond	level conditioned (threshold condition)

Acknowledgements

The Linguistic Atlas of Japan Database (LAJDB) was supported by a Grant-in-Aid for Publication of Scientific Research Results (database) in 2001, 2002, 2003, 2004, 2005, and 2008 (Project Leader: Yasuo Kumagai). This paper includes some outcomes of the collaborative research projects "Analyzing large-scale dialectal survey data from multiple perspectives" (2009–2012, Project Leader: Yasuo Kumagai), "General Research for the Study and Conservation of Endangered Dialects in Japan" (2013–, Project Leader: Nobuko Kibe) at the National Institute for Japanese Language and Linguistics, and "Development of quantitative methods of analyzing large-scale dialectal distribution data" (Grants-in-aid for scientific research (c), JSPS KAKENHI Grant Number 26370555, 2014–, Project Leader: Yasuo Kumagai). I thank the anonymous reviewers and the editors for their valuable comments and suggestions.

References

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge, New York: Cambridge University Press.

Hondô, Hiroshi. 1980. Gendai hyôjun nihongo no bunpu: Nihon gengo chizu de mite [Distribution of modern standard Japanese: An observation by using the LAJ]. In Shigeru Satô (ed.), *Sato shigeru Kyoju taikan kinen ronshu kokugogaku.* 479–498. Tokyo: Ohfusha.

Ichii, Tokiko. 1993. *Hôgen to keiryô bunseki [Dialect and quantitative analysis].* Tokyo: Shintensha.

Inagaki, Shigeko. 1980. Hôgen sesshoku to gokei heiyô: "Nihon gengo chizu" no bunpu kara [Dialect contact and doublets: Some examples from the distributions in the LAJ]. *Tôkyô toritsu daigaku hôgen kenkyûkai kaihô* 92. 1–10.

Inoue, Fumio. 2001. *Keiryôteki hôgen kukaku [Quantitative dialect division].* Tokyo: Meiji shoin.

Inoue, Fumio. 2004. Heiyô genshô to gengo genshô no chûkan dankai: Kasai data 3 kurasutâ no hukyû katê [Joint usage of forms and intermediate stages of linguistic change: Process of diffusion of 3 clusters of Kasai data]. *Gogaku kenkyûjo ronshû (Journal of the Institute of Language Research)* 9. 1–19.

Kasai, Hisako. 1981. Hyôjun gokei no zenkoku bunpu [Nationwide distribution of standard forms]. *Gengo seikatsu* 354. 52–54.

Kobayashi, Takashi. 2004. *Hôgengakuteki nihongoshi no hôhô [Method of dialectological study of Japanese language history]*. Tokyo: Hituzi Syobo.

Kokuritsu Kokugo Kenkyûjo (NLRI). 1966. Nihon gengo chizu kaisetsu: Hôhô [Introduction to the linguistic atlas of Japan: Methodology]. In Kokuritsu Kokugo Kenkyûjo (NLRI) (ed.), *Nihon gengo chizu (Linguistic atlas of Japan)*. Tokyo: Printing bureau, Ministry of Finance.

Kokuritsu Kokugo Kenkyûjo (NLRI). 1966–1974. *Nihon gengo chizu (Linguistic atlas of Japan)*. Tokyo: Printing bureau, Ministry of Finance.

Kumagai, Yasuo (ed.). 2013a. *Daikibo hôgen dêta no takakuteki bunseki seika hôkokusho: Gengo chizu to hôgen danwa shiryo [Analyzing large-scale dialectal survey data from multiple perspectives]*. Tokyo: Kokuritsu Kokugo Kenkyûjo (NINJAL).

Kumagai, Yasuo. 2013b. Development of a way to visualize and observe linguistic similarities on a linguistic atlas. *Working papers from NWAV Asia-Pacific* 2.

Kumagai, Yasuo. 2013c. Nihon gengo chizu no dêtabêsuka to keiryôteki bunseki: Heiyô genshô, hyôjungokei no bunpu to kôtsûmô, hôgenruijido [Development of database of Linguistic atlas of Japan and quantitative analysis]. In Yasuo Kumagai (ed.), *Daikibo hôgen dêta no takakuteki bunseki seika hôkokusho: Gengo chizu to hôgen danwa shiryo [Analyzing large-scale dialectal survey data from multiple perspectives]*, 111–128. Tokyo: Kokuritsu Kokugo Kenkyûjo (NINJAL).

Satô, Ryôichi. 1986. Chiikishakai no kyôtsûgoka [Standardization in regional society]. In Kiichi Iitoyo (ed.), *Hôgenkenkyû no mondai*, vol. Kôza hôgengaku, 145–178. Tokyo: Kokusyokankôkai.

Sibata, Takesi & Yasuo Kumagai. 1993. The S&K Network method: Processing procedures for dividing dialect areas. *Zeitschrift für Dialectologie und Linguistik* 74. 458–495.

Takada, Makoto. 1969. Kotoba no chiri: Nihon gengo chizu kara [Geography of words, Kyûshû district: An observation by using the LAJ]. *Gengo seikatsu* 216. 30–38.

Tanaka, Akio. 1991. *Hyôjungo: Kotoba no komichi [Standard language: A lane of speech]*. Tokyo: Seibundô Shinkôsha.

Tokugawa, Munemasa. 1993. *Hôgenchirigaku no tenkai [Development of dialect geography]*. Tokyo: Hituzi Syobo.