Development of a Way to Visualize and Observe Linguistic Similarities on a Linguistic Atlas*

Yasuo Kumagai
National Institute for Japanese Language and Linguistics

1. Introduction

Dividing dialect areas has been one of the fundamental issues in Japanese dialectology from its beginning. In mid 1980s, we started developing S&K Network Method as a method for dividing dialect areas quantitatively through a network-based approach. In our earliest attempt, we had two aims: (1) to objectively divide geographical space into dialect areas on the basis of linguistic features; (2) to observe varying degrees of similarity among localities and how these localities form clusters on a map. With the development of this method, a greater focus has been laid on the latter aim. In discussing dialect areas, the classification of dialects and the division of a dialect continuum have sometimes been confused with each other. The S&K Network Method is primarily oriented to the division of dialect continua, not to the classification of dialects.

2. Data

We are dealing with dialectal data acquired by linguistic geographical surveys. It is after the maps are drawn and interpreted that the input data for the Network Method are prepared. The features which show significant distribution on a map are selected as the items of the input data. Thus, the input data are the result of interpretation and analysis.

Our method was developed by mainly using Linguistic Atlas of Amami-oshima (LAA) and *Linguistic Atlas of Japan* (LAJ) vols.1-6. The survey of LAA was carried out by Sibata et al. in 1977-79. The total survey points were 143. LAA data was the first data we used for the development of our method and was composed of 177 phonological items (LAA-ph) and 138 lexical items (LAA-lx).

As the next step, to test the applicability to larger data and to work with a nationwide dialect data, we began to apply the Network Method to LAJ, which had 2400 survey points, in 1996. The survey of LAJ was carried out from 1957 to 1965 by The National Language Research Institute, NLRI, the institute preceding the present NINJAL. The number of questionnaire items (Q items) were 285 (mainly lexical field). The data of LAJ vol. 3 was

computerized around 1985 by the members of NLRI. As the input data, 27 Q items which had almost all survey points were selected and number of the input data items was 4151. From 1999, we have been constructing *Linguistic Atlas of Japan Database*, LAJDB[1], to make all the information contained in LAJ available on computer (Kumagai 2007). At present, 119 Q items have been completed. In this paper, 55 Q items which has almost all survey points are selected from LAJDB for the input data. The number of the input data items is 8611.

3. Network representation NT-1(r)n

The Network Method is a general term for a series of methods which we have developed. Here, we will glance at NT-1(r) type, that is, NT-1(r)n and NT-1(r)d (Sibata&Kumagai 1985,87,93). In NT-1(r)n, the degree of similarity between any two localities is measured by the number of linguistic features shared by the two localities. We call this similarity matrix NC (fig.1 (1)). In NT-1(r)d, the measure of linguistic similarity between two localities is the degree of similarity between the distribution patterns of NC they have. That is, the degree of similarity between two localities is measured by the degree of similarity between the relationship patterns which each locality shows in relation to all of the localities. We use Euclid distance as the mesure. The matrix obtained is a distance matrix and we call this DC.

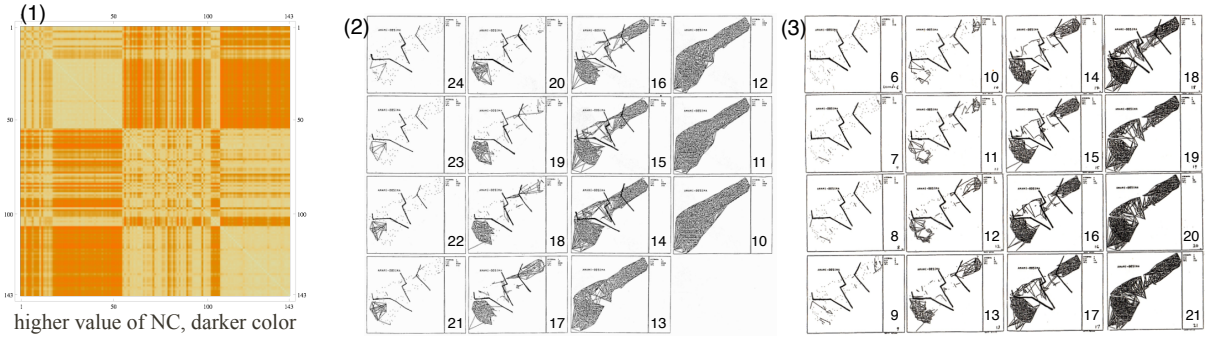

Figure 1 LAA-ph, (1) visual display of similarity matirix NC (143*143) / (2) Network Representation of NT-1(r)n (Lcond=24~10) / (3) NT-1(r)d (Lcond=6~21, Value of DC categorized within the range of 100) / ((2) and (3) are from Sibata & Kumagai(1993). Isogloss bundles by T.Sibata (bold lines) was superposed on the NT-(r)n and d in (2) and (3).)

NT-1(r) is a very simple way to visualize and observe NC and DC, that is, a matrix of linguistic similarities and dissimilarities among every pair of points. NT-1(r) draws a line between two any points when the similarity between them satisfies the threshold condition, Lcond, which is taken as variable to see the patterns of each level of similarities and the change of the patterns of the network varying with the change of Lcond (fig.1 (2), (3)).

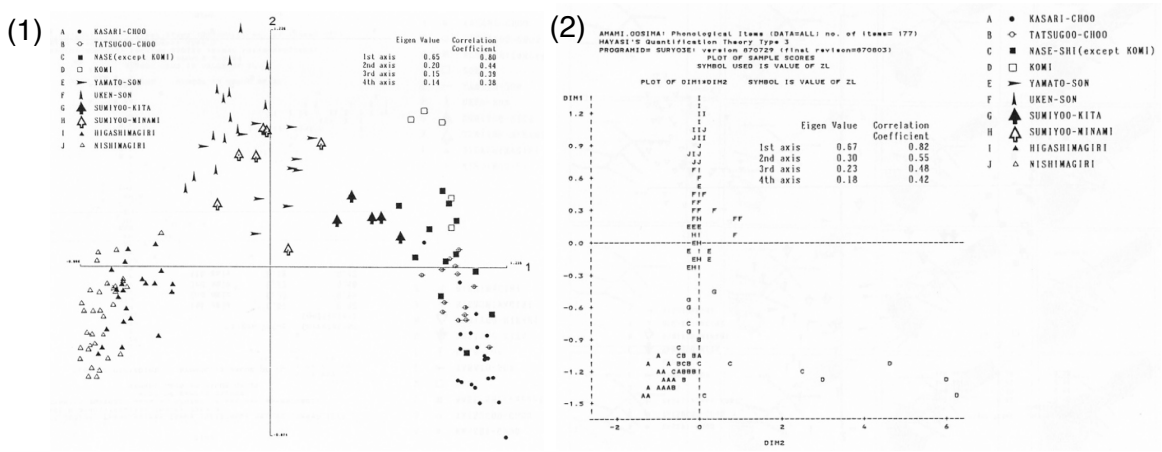

Figure 2 Plot of scores of Hayashi's type 3 given to localities (1st * 2nd axis), LAA-ph
(1) Items of less than 6 responses were omitted, (2) All items. (from Sibata & Kumagai 1993)

We compared the S&K Network Method with other approaches (Sibata&Kumagai 1993 et.al). Hayashi's quantification theory type 3 (Hayashi's type 3) is a kind of multivariate analysis and is widely used in Japan. It corresponds to the correspondence analysis. It sorts or classifies the items and objects at the same time. Inoue (1986) applied Hayashi's type 3 to divide dialect areas and argues that the method is quite suitable for processing nominal scale data, which is the same type as ours. But, it is necessary to select items to get clear patterns. Since the items which only have few responses tend to distort the patterns obtained, calculating all the items as a whole is often not revealing. Some image of this nature can be obtained by comparing fig.2 (1) and (2). Hayashi's type 4 is a kind of MDS. A similar problem occurs, in this case, too. The reason is that these multivariate analyses are optimized as regards to separating or distinguishing the groups of the objects in multidimensional space. The NT-1(r) representations enable us to observe the data, which shows the relations among the localities, in a straightforward way. Our data are from linguistic geographical surveys and have significant geographic distribution patterns. Because of this nature, NT-1(r) works well.

To grasp the nature of the distribution patterns sorted and classified by Hayashi's type 3, we made a simple graphical presentation. We arranged the distribution pattern of each item and the plots of the Hayashi's Type 3 scores corresponding to it on a picture. We made the series of these pictures sorted by the values of an axis of Hayashi's type 3 and by a result of cluster analysis as regard to an output of Hayashi's type 3 and made the animations of these series of the pictures so as to help observing the nature of distribution patterns sorted by Hayashi's type 3. Fig.3 shows some samples. The scores of fig.3 (1), (2), (3) are very close and can be classified as the same distribution pattern.

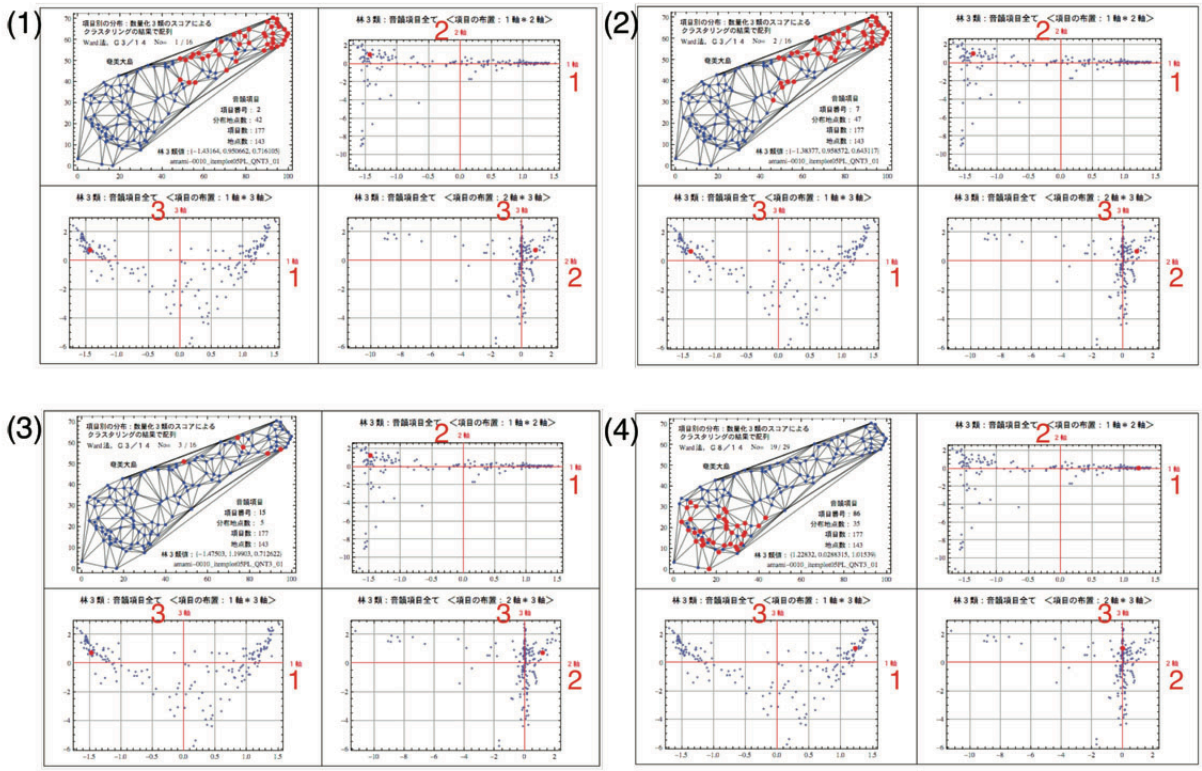

Figure 3 Sample pictures of a series of distribution pattern grouped by a result of cluster analysis (Ward method) as regard to an output of Hayashi's type 3. ((1), (2), (3) are grouped as the same distribution pattern. (4) belongs to a different group.)

4. Network representation on the Delaunay net

As a way to represent continuity among survey points on the geographical space in a formal manner, we used Delaunay triangulation as an approximation (Kumagai 1996b, 2002 etc). Delaunay triangulation is a computational geometrical method which enables us to get adjacent points from randomly distributed points on a plane. Delaunay triangulation is a widely used method in many scientific areas. It gives us mathematically defined adjacent points. Here, we must remember that this continuity represented by Delaunay triangulation is a kind of an approximation or a tool for our analysis.

We assign a value of NC to a line which connects two adjacent points, to visualize how linguistically similar survey points are distributed on a map, that is, how linguistic similarities between adjacent points change over a map. We call a network of the points made by Delaunay triangulation a Delaunay net and we call the representation of NC values on the Delaunay net Network representation on the Delaunay net. We applied Delaunay triangulation to LAJ and Amami-oshima data (fig.4). We can observe a picture of continuity of linguistic similarities on a map.

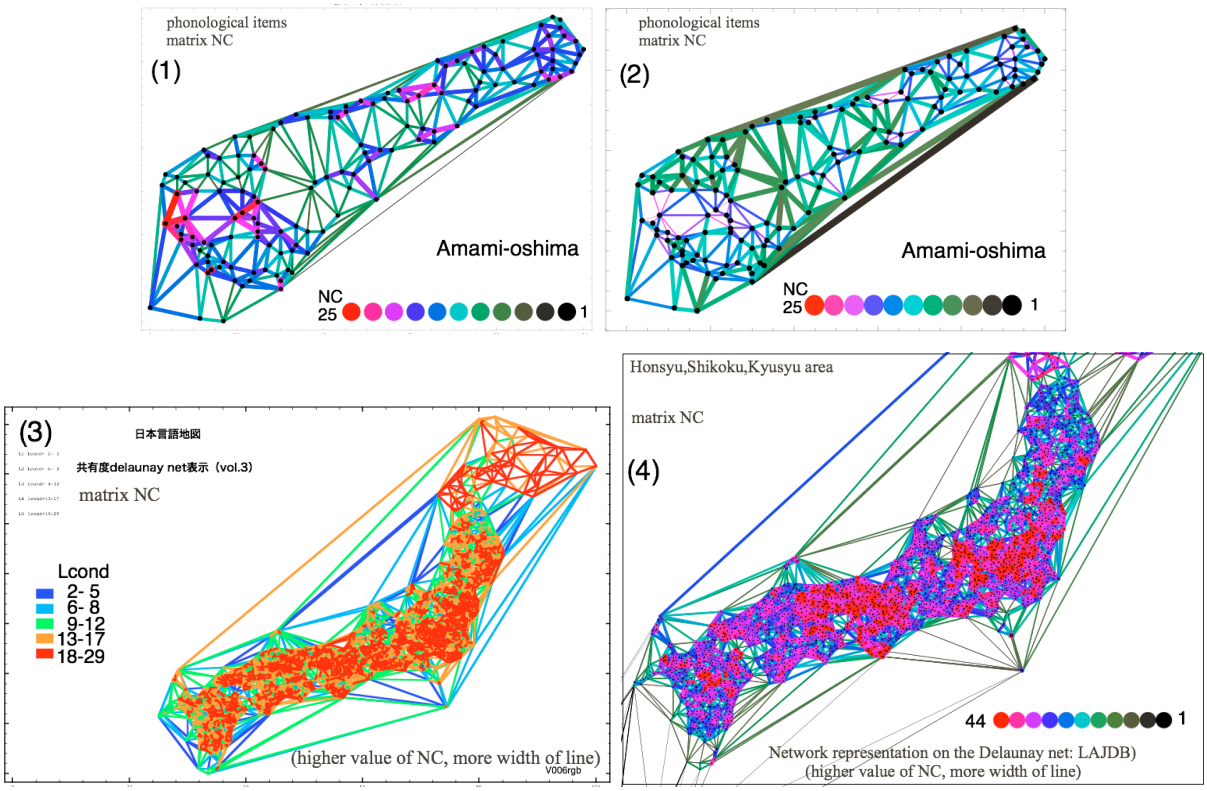

Figure 4　Network representation on the Delaunay net.　(1)LAA-ph (from Kumagai 2008) / (2) LAA-ph, type2 representation (from Kumagai 2008) / (3) LAJ vol.3 (from Kumagai 1996b) / (4) LAJDB

[Higher value of NC, more width of line. In type 2, Higher value of NC, less width of line]

5. An example of application of network analysis (minimum spanning tree)

Now we have a formal and approximate representation of continuity among the survey points and representation of the values of similarity between each of the pairs. These elements corresponds to a network and values of the lines of the network. In network theory or graph theory, a network with values or weights of lines is called a weighted graph (network). We can apply a network analysis to this network (Kumagai 2008).

Minimum spanning tree (MST) is a network which has a tree structure. That is, MST is the set of weighted lines selected from the lines of original network and total weight of the network will be minimum and it covers or connects all of the points. Here we can take a weight as a cost. We calculated MST as regards to the Delaunay net representation of LAA-ph data (Kumagai 2008) and of LAJDB data. We set the costs (weights) by means of a similarity matrix NC so that the less the similarity, the more the cost. To find a MST is to find out a tree of route on a map avoiding high cost route, where linguistic similarity NC is low

and the total cost is minimum. This is a kind of optimization on a network. We can observe that the MST appear on the continuity of linguistically similar points (fig.5).

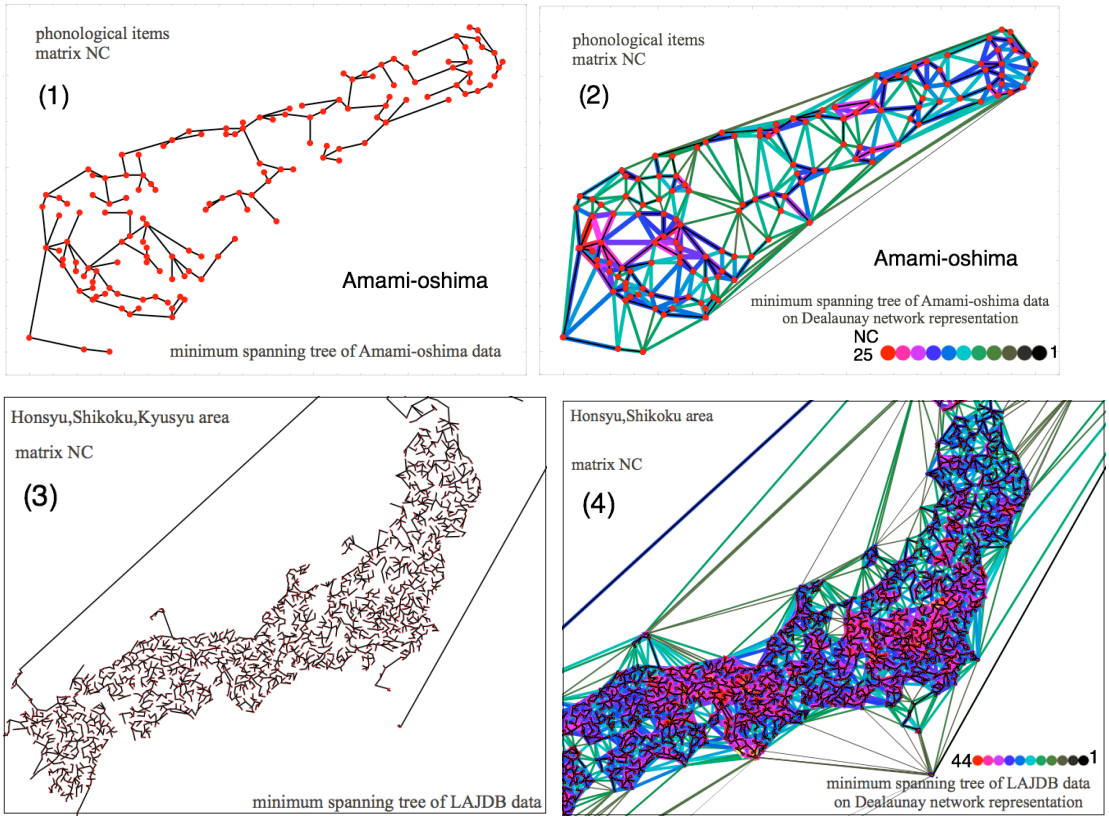

Figure 5 (1) MST of LAA-ph (from Kumagai 2008) / (2) MST of LAA-ph on Dealaunay network representation (from Kumagai 2008) / (3) MST of LAJDB data / (4) MST of LAJDB data on Dealaunay network representation

6. Network representation NT-1(r)n on Delaunay net

We imposed NT-1(r)n representation on Delaunay net representation (fig.6). Delaunay net representation makes it possible to observe the transition or change of the linguistic similarities of adjacent points but it concerns no other points except the neighboring points. NT-1(r) enables us to see relationships of similarity not restricted to the ones between adjacent points. By overlaying these two types of representation, we can observe the distribution of similarities along the continuity and the one on all over the map, that is not restricted to neighbors, at the same time. In transitional zone and homogeneous zone, Nt-1(r) shows us different network structures. So, by overlaying the two kinds of representation, we can distinguish two types of distribution patterns of similarities which the network representation of Delaunay network can not distinguish. It is useful to combine more than one method. Each method has its own merit, demerit and its own perspective.

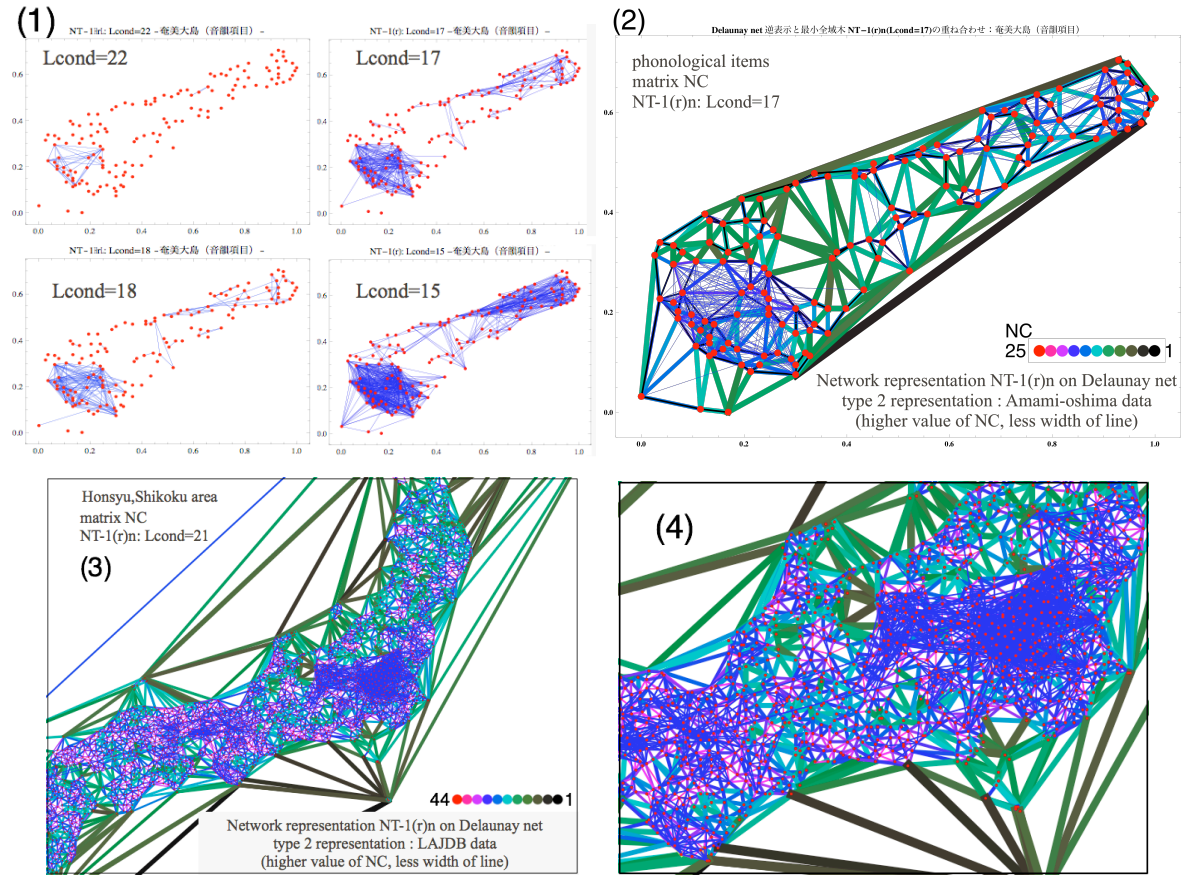

Figure 6 (1) A Network Representation of NT-1(r)n: LAA-ph (from Kumagai 2008) / (2) Network representation NT-1(r)n on Delaunay net type 2 representation : LAA-ph (from Kumagai 2008) / (3) Network representation NT-1(r)n on Delaunay net type 2 representation: LAJDB / (4) Enlarged partial detail of (3)

7. Concluding remarks

The concept of network is a useful key in our development and has wide implications. Thinking network is also thinking contact at the same time. Contact is also a key concept. Distributions of dialects, speech communities and dialect areas are all networks and they are developed from the contacts of persons. Now, our objectives are (1) to extract latent information and structure of the data appropriately to the nature of the data itself, (2) to develop a new method to visualize the dynamics, flows and trends of dialectal distribution based on the data, (3) to understand the distribution pattern of dialect in relation to the dynamics of language. Our first aim of dividing dialect areas can be placed in such context.

References

Inoue, Fumio. (1986). Bunpo gensyo ni yoru keiryoteki hogen kukaku (Quantitative dialect division by means of grammatical phenomena). Gengo Kenkyu 89, 68-101.

Kumagai, Yasuo. (1996a). Nettowakuho ni yoru Nihon gengochizu dai 3 syu no chiten-kan ruijido no sokutei to deta no shikakuka: tairyo deta e no tekiyo no kokoromi (Measurement of the linguistic similarities among the survey points of Linguistic Atlas of Japan Vol.3 and the visualization: an application to the large amount of data). Conference papers of the dialectological circle of Japan 63. 31-40.

_____. (1996b). Nihon gengochizu no Delaunay net zyo ni okeru rinsetsu chiken-kan kyoyudo no hyozi (Representation of linguistic similarities of adjacent survey points of Linguistc Atlas of Japan on Delaunay net). Additional handout of (Kumagai 1996a).

_____. (2002). Keiryoteki hogen kukaku to hogen chirigaku: keiryoteki hogen kukaku no tame no nettowakuho no kaihatsu o tosite (Quantitative division of dialect area and dialect geography: throught development of Network method for quantitatively dividing dialect area). In Yoshio Mase (ed), Hogen chirigaku no kadai. Tokyo: Meiji shoin. 150-154.

_____. (2007). Nihon gengochizu no Detabesu-ka (The Linguistic Atlas of Japan Database). Conference papers of the dialectological circle of Japan 85. 27-34.

_____. (2008). Gengo chizu-jo no chiten-kan no ruiji kankei no shikakuka to bunseki: Nettowakuho ni okeru Gurafu riron no oyo to shimyuresyon no kokoromi (Visualization and analysis of the relationships among survey points on the linguistic atlases: An application of graph theory to the S&K Network-method and an experiment of computer simulation). Conference papers of the dialectological circle of Japan 87. 43-52.

Sibata, Takesi and Yasuo Kumagai. (1985). Gengoteki tokucho ni yoru chiiki bunkatsu no tame no "nettowakuho": tokuni NT-1(r) ni tsuite (The Network Method: a method for divideing an area on the basis of linguistic features: with special reference to NT-1(r). Kokugogaku 140. left 45-60.

_____. (1987). Nettowakuho ni okeru chitenkan no gengoteki ruiji no atarashii toraekata to syori no sikata: gengoteki tokucho ni yoru chiiki bunkatsu no tameno Nettowakuho 2 (A new "Network method" and its processing procedures for dividing dialect areas). Kokugogaku 150. left1-14.

_____. (1993). The S&K Network Method: Processing Procedures for Dividing Dialect Areas. Zeitschrift für Dialectologie und Linguistik 74. 458-495.

• This presentation includes some outcomes of the collaborative research project “Analyzing large-scale dialectal survey data from multiple perspectives” (2009-2012) at NINJAL.

[1] Linguistic Atlas of Japan Database (LAJDB) was supported by Grant-in-Aid for Publication of Scientific Research Results –Database - in 2001,2002,2003,2004,2005,2008.