東京外国語大学論集　52　（別刷）

COMPUTATIONAL DIALECTOLOGY (1)

Fumio INOUE

1996

Area and Culture Studies Vol. 52(1996)

COMPUTATIONAL DIALECTOLOGY (1)

Fumio INOUE

0.1. INTRODUCTION

The International Congress is a convenient occasion for scholars who are working in the same field to communicate. The many meetings of "Methods in Dialectology" and the First International Congress of Dialectologists in Bamberg have been good meeting points for dialectologists who deal with dialectal data by computer.

For this subject, the first volume of the "Proceedings of the First International Congress of Dialectologists" in Bamberg is an important reference book. The book itself is a product of desktop computer, making full use of the efficiency of a computer. It includes a convenient name index and subject index. The papers included in the Proceedings will be mentioned in appropriate places in my lecture.

I have tried to summarize recent advances in computational dialectology, including as many studies as possible by various scholars in the world, but because of time limit I had to concentrate on Japanese computational dialectology.

The computational approach is also flourishing in the field of social dialectology, that is, the study of dialect usage in a community. To my great regret, these related fields will not be discussed here because of the time limit.

0.2. SUMMARY

In this lecture I would like to advocate the usefulness of computers in dialectology. In the first and second sections, I will survey simple methods using arithmetic calculation, in the third section more complicated methods of multivariate analyses, and in the fourth section the simple but effective computational method the "gravity center method". I will also discuss that both geographical and linguistic information should be subjected to computer analysis. Distinction of numerical data and non-numerical (nominal) data is also important here for computational procedure.

Computational study of dialects seems to be flourishing all over the world now. Many atlases drawn up by computer have appeared and several original statistical methods have been applied in many parts of the world. Computational analysis is also flourishing in the far eastern tip of Eurasia. Some atlases have been compiled by computer and several original statistical methods have been applied in Japan.

1. MAP DRAWING BY COMPUTER

The most important and time-consuming work for linguistic geography after the collection of raw data is the drawing up of maps. The application of computation is useful first in drawing clear maps. Recent computer produced maps are more clear and precise than

maps drawn by the hands of dialectologists. Several mapping programs are now being used on computers and many linguistic atlases have been published. Computerized linguistic atlases have appeared in many parts of Europe. Marburg seems to be the center of computational dialectology. As shown in **Figure 1-1**, "Survey of English Dialects", which

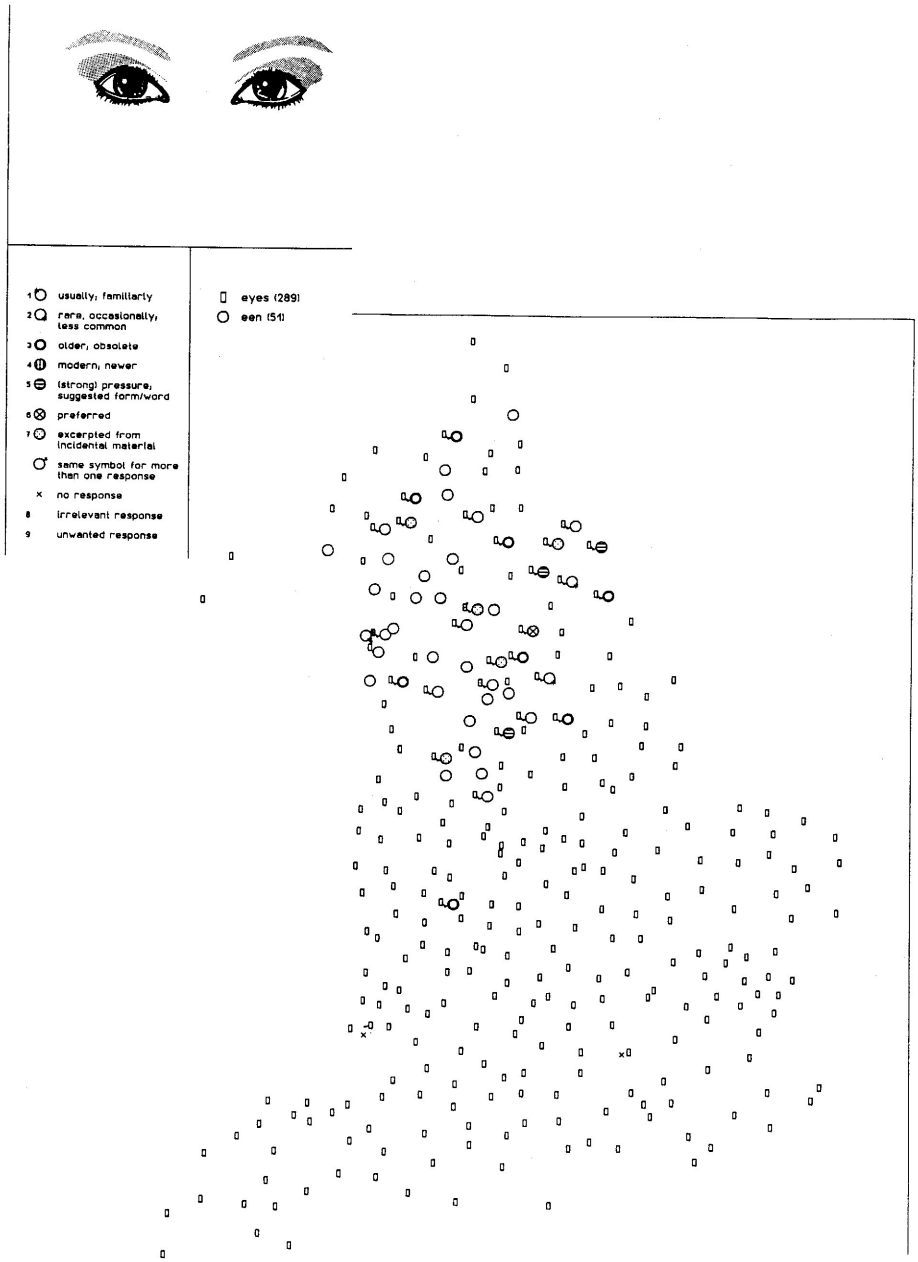


Figure 1-1 "Computer Produced Linguistic Atlas of England" by Viereck (1992) based on the "Survey of English Dialects"

was originally published as raw transcription data in several books, has been published in the form of a linguistic atlas by Prof. Viereck (1992a). The large amount of data gathered in the past is now available in map form thanks to the computer.

The first attempt at computerized mapping in Japan was made early in the 1960's but only one trial map was made (Tokugawa 1993). This was attempted as a joint project by a dialectologist and a specialist in computation who was interested in the possibility of computers.

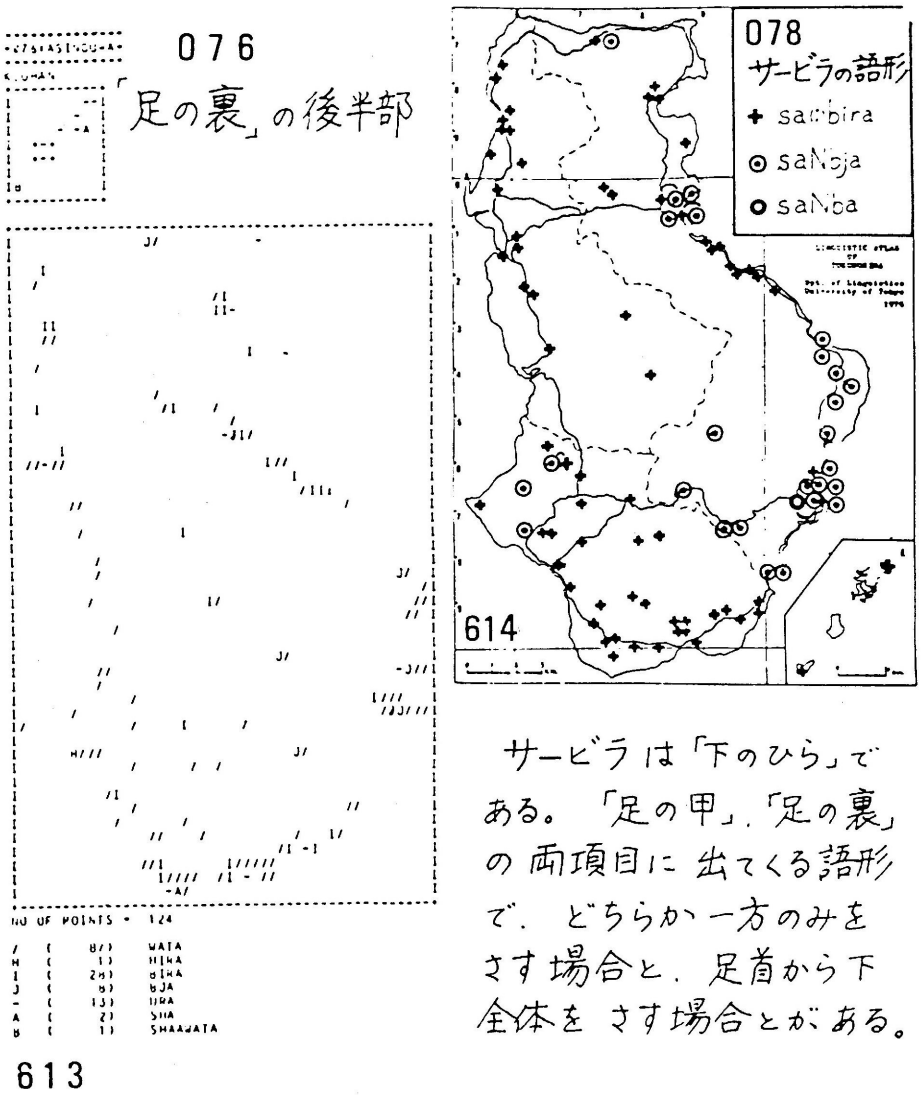


Figure 1-2 "Linguistic Atlas of Tokunoshima" by students of the University of Tokyo under the guidance of Prof. Takesi Sibata.

As for Japan, Tsunao Ogino's GLAPS, a package program for mapping and for statistical analysis, was produced in the 1970's (1994). Some scholars, including myself, have used this program and several linguistic atlases based on this have been drawn up by computer. **Figure 1-2** is an example from the "Linguistic Atlas of Tokunoshima", which was produced by students of the University of Tokyo under the guidance of Professor Takesi Sibata (1977). You can compare this with a hand-made map.

The advantage of using computers in drawing up maps is the clarity of the figures which make a photo-ready draft possible. **Figure 1-3** is a sample from some recent publications using a desktop computer, by Kenji Takahashi (1991).

When the research method is well designed for computer mapping, maps can be produced quickly. When T. Ogino and I planned to gather data of new dialect forms used

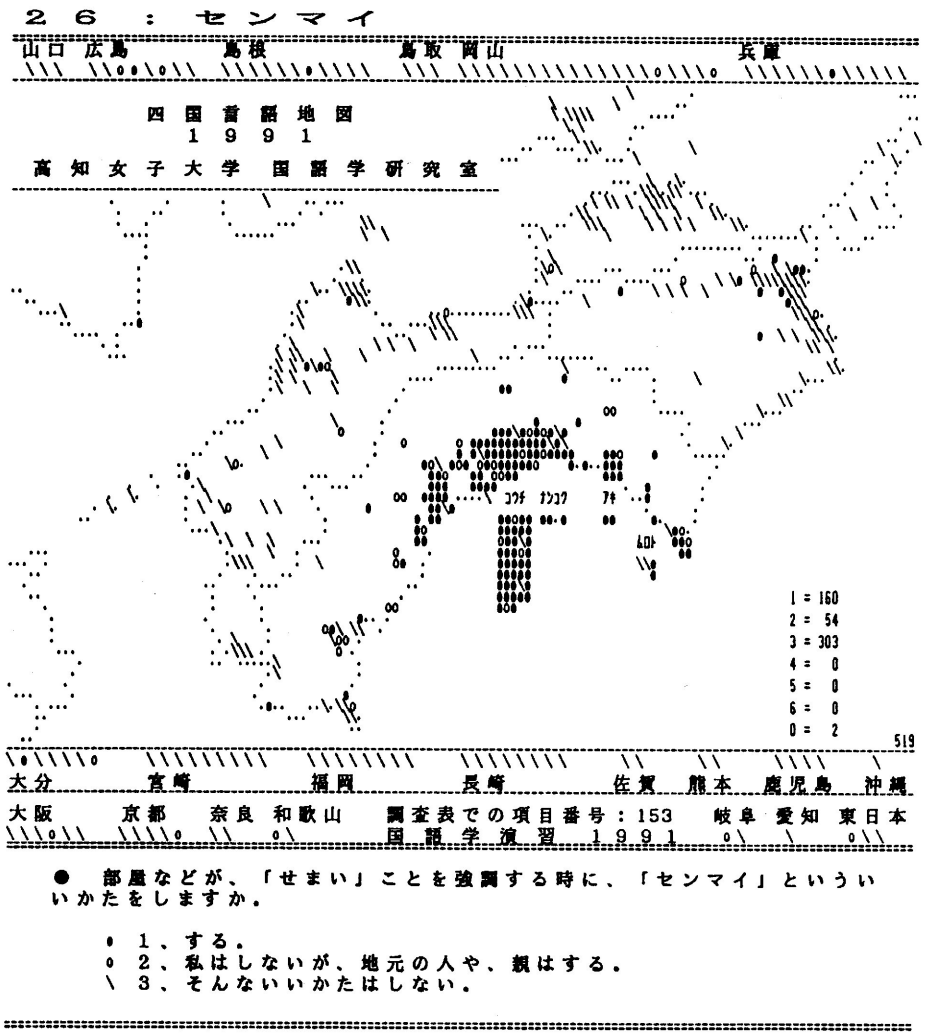


Figure 1-3 "Linguistic Atlas of Shikoku" by Takahashi (1991)

by junior high school students of all over Japan (Inoue & Ogino 1984), anticipated word-forms had been entered in the questionnaire with code numbers, and students were asked to tick the word-forms they use or hear (**Figure 1-4**). The distribution maps were made within several weeks, and an atlas was published within several months.

But the same work can still be done by human hand more meticulously, and thus by merely drawing up maps we do not make full use of the potential of computers. The main obstacle for computational mapping is the large effort necessary for data input when the word-forms have not been coded beforehand. If we sincerely consider "cost-perfor-

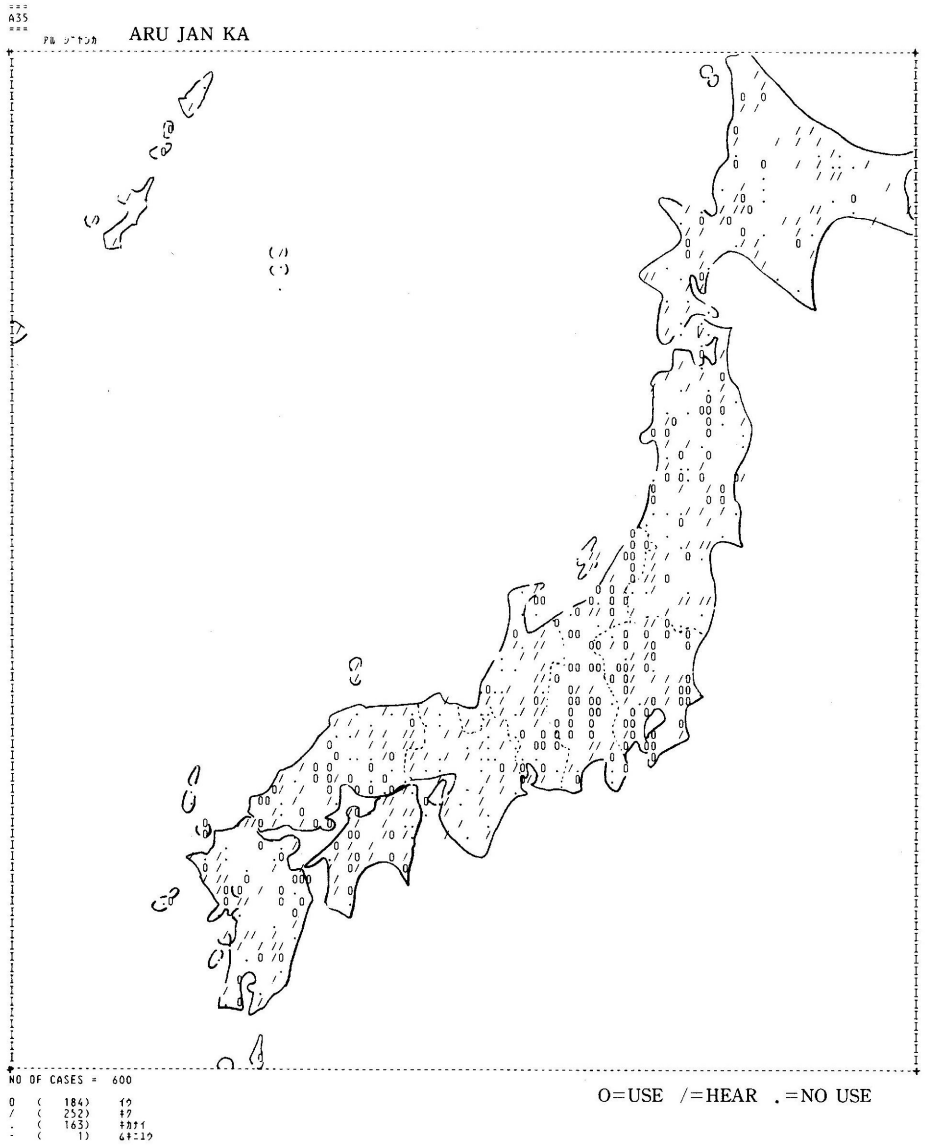


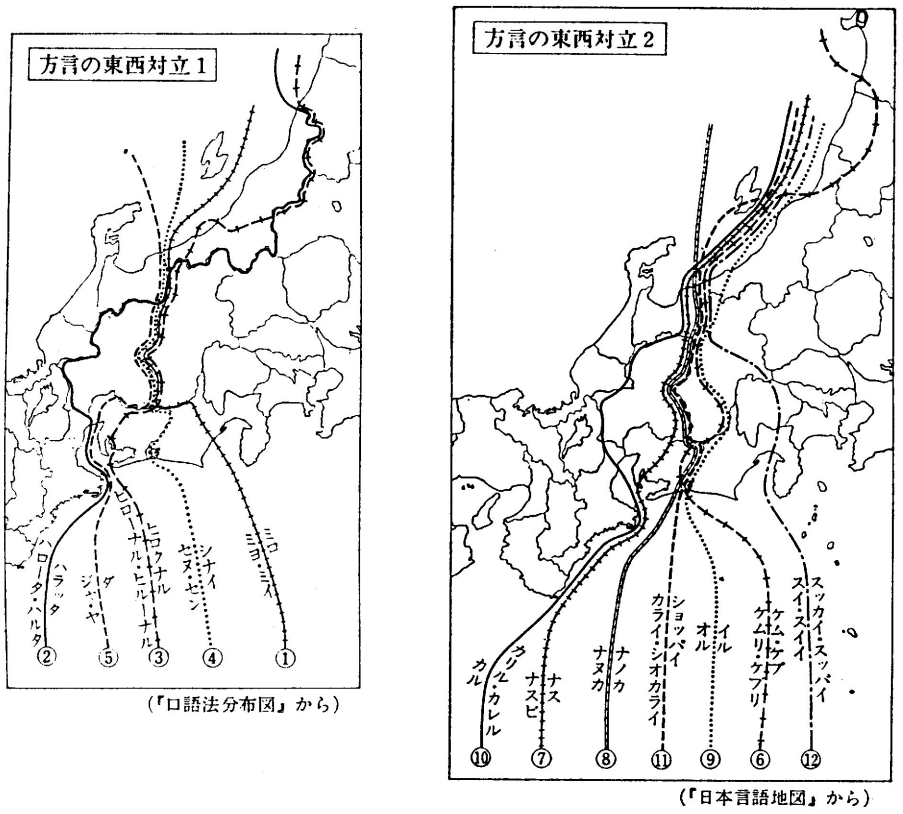
Figure 1-4 "Neologisms in Japanese: Materials" by Ogino and Inoue (1984)

mance", using a computer to make a single map is not profitable. Drawing up a map by hand using rubber stamps or "letraset" is more time-saving. Quantitative processing of data, especially dialect division (or classification) is the field which receives most benefit from the use of computer.

2. QUANTITATIVE STUDY——ARITHMETIC METHODS

Computational methods become more effective if the data for mapping is also processed quantitatively. Quantitative dialectology or dialectometry is a promising field for computers.

Three main methods for quantitative dialectology will be discussed here: the ISOGLOSS METHOD, IDENTITY METHOD and GRAVITY CENTER METHOD. All three methods are arithmetic because they can be executed by hand if one has a calculator and enough time. But multivariate analyses, which will be discussed in the next section, cannot be executed without a computer. The third arithmetic method, the gravity center method, will be discussed after the section on multivariate analysis because the results of the gravity center method can be better interpreted after an application of multivariate



(『口語法分布図』から)

(『日本言語地図』から)

Figure 2-1 Bundles of Isoglosses in Central Japan by Tokugawa (1981) based on the "Linguistic Atlas of Japan"

analyses and because the gravity center method is effective for summarizing the results of multivariate analyses.

2.1. ISOGLOSS METHOD

The first arithmetic method of quantification is the ISOGLOSS METHOD. An ISOGLOSS or heterogloss is the classical means by which differences of dialects were shown in traditional dialectology. Sometimes linguistic maps of individual items themselves have been drawn up by means of isoglosses, as in Germany and in England.

Bundles of isoglosses are often sought for in order to find dialect boundaries. Earlier studies in dialectology in Japan at the beginning of 20th century paid great attention to the dialectal division of Japanese. The exact location of the border between eastern and western dialects in central Japan has been an especially controversial issue. The locations of bundles of isoglosses were later more clearly shown by Tokugawa (1993) without the

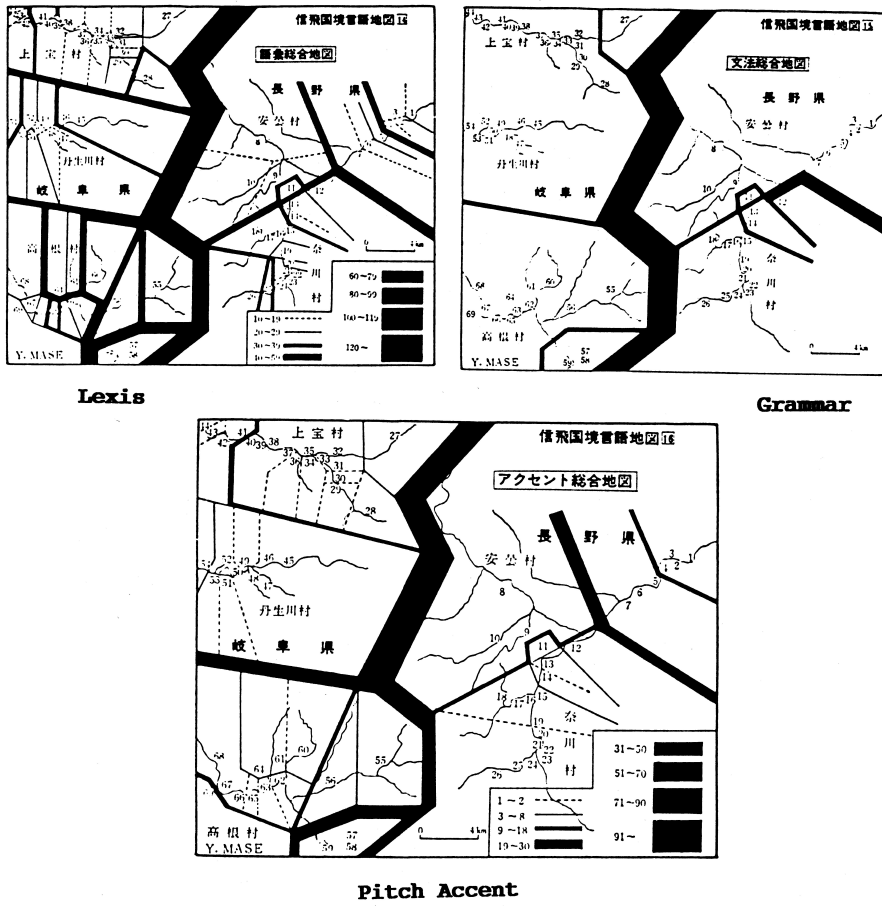


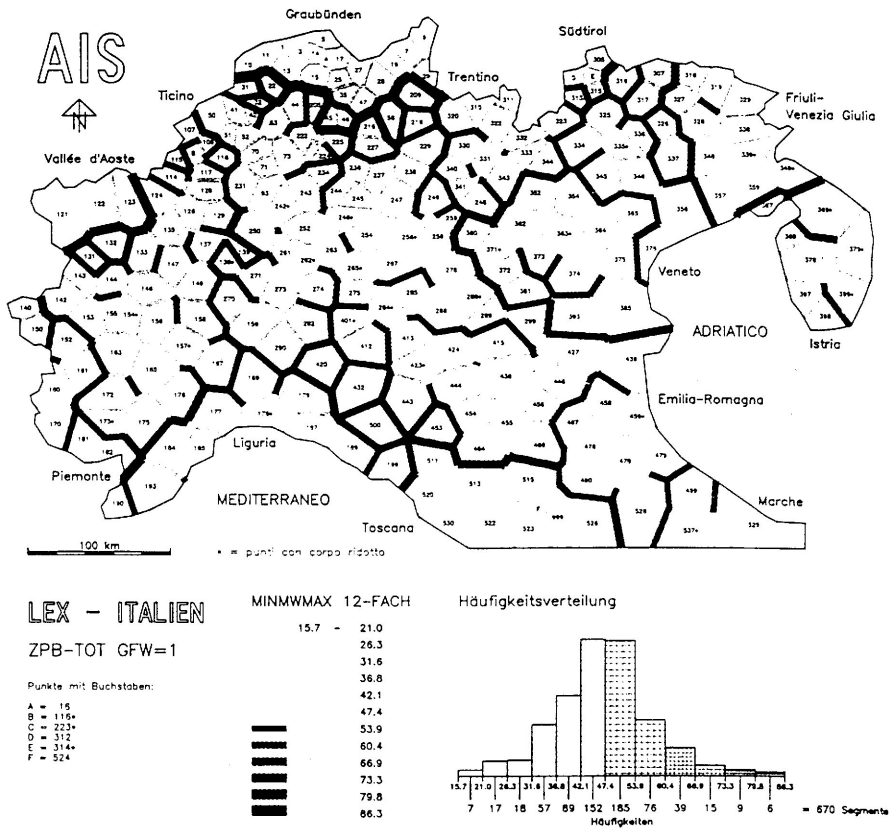
Figure 2-2 The HONEYCOMB METHOD applied to Japanese Dialects in central Japan by Mase (1964)
Lexis, Grammar and Pitch Accent

use of a computer. **Figure 2-1** shows bundles of isoglosses in central Japan. This is an attempt at finding dialect boundaries by means of the cumulative isogloss method.

There are several problems with the isogloss method. The number of items (words) counted in these studies were usually not (nearly) enough, reflecting the limitations of human effort. Moreover selection of phenomena for drawing isogloss was sometimes ad hoc and arbitrary.

A newer technique which avoids the problem of arbitrariness is the HONEYCOMB METHOD applied by Grosse (1955) for a German dialect area. Séguy's analysis (1973) of French dialects is similar in principle. As shown in **Figure 2-2**, this method was applied to Japanese dialects by Mase (1964) without the use of a computer in a mountainous area in central Japan where many important isoglosses dividing eastern and western Japan run.

In the ideal honeycomb method, all the phenomena in all the area investigated are taken into consideration. Thus elimination of arbitrary, subjective judgment is possible. Sometimes weighting of phenomena is attempted by attaching various degrees of importance



Karte 7: Wabenkarte (auch: Schotten-oder Grenzsegmentkarte)
Intervallalgorithmus: MINMWMAX

Figure 2.3 The HONEYCOMB METHOD in Northern Italy by Goebl (1993)

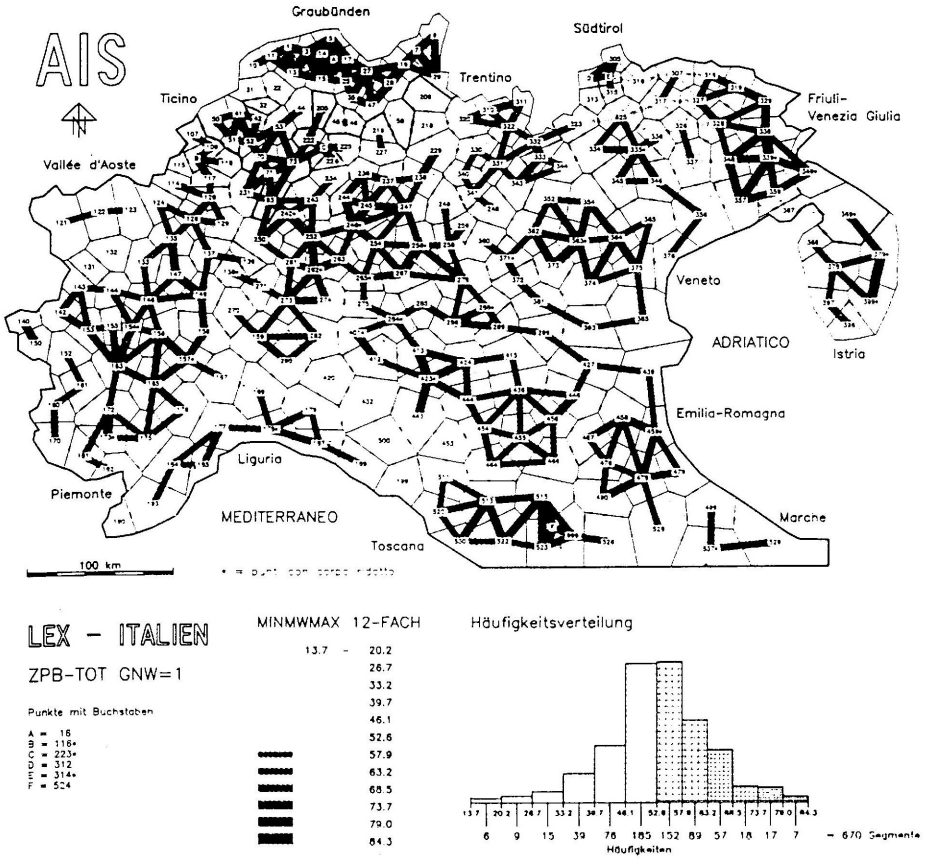
according to the frequency of usage or linguistic system.

Figure 2-3 shows Goebl's recent application of the honeycomb method by computer (1993). Thick lines suggest borders of dialects. In **Figure 2-4**, he further applied a reverse technique called "Strahlenkarte" or "carte à rayons". In this case it is not the borders, but close dialectal similarities that are shown by thick lines.

It is noteworthy here that in both techniques only neighboring localities are considered. However the selection of neighboring localities is itself sometimes problematical.

There still remains the fundamental problem of how to draw isoglosses when word-forms show scattered distribution. This often happens in areas near dialectal borders, not to mention the scattered distribution because of language standardization. In these cases isoglosses are difficult to draw. As this problem is inevitable in American dialects, Linn and Regal's method (1993), shown in **Figure 2-5**, of contour maps is a good solution to this problem. But it is difficult to apply this technique to the honeycomb method.

The separate distribution of the same phenomena over a large scale is often found in the

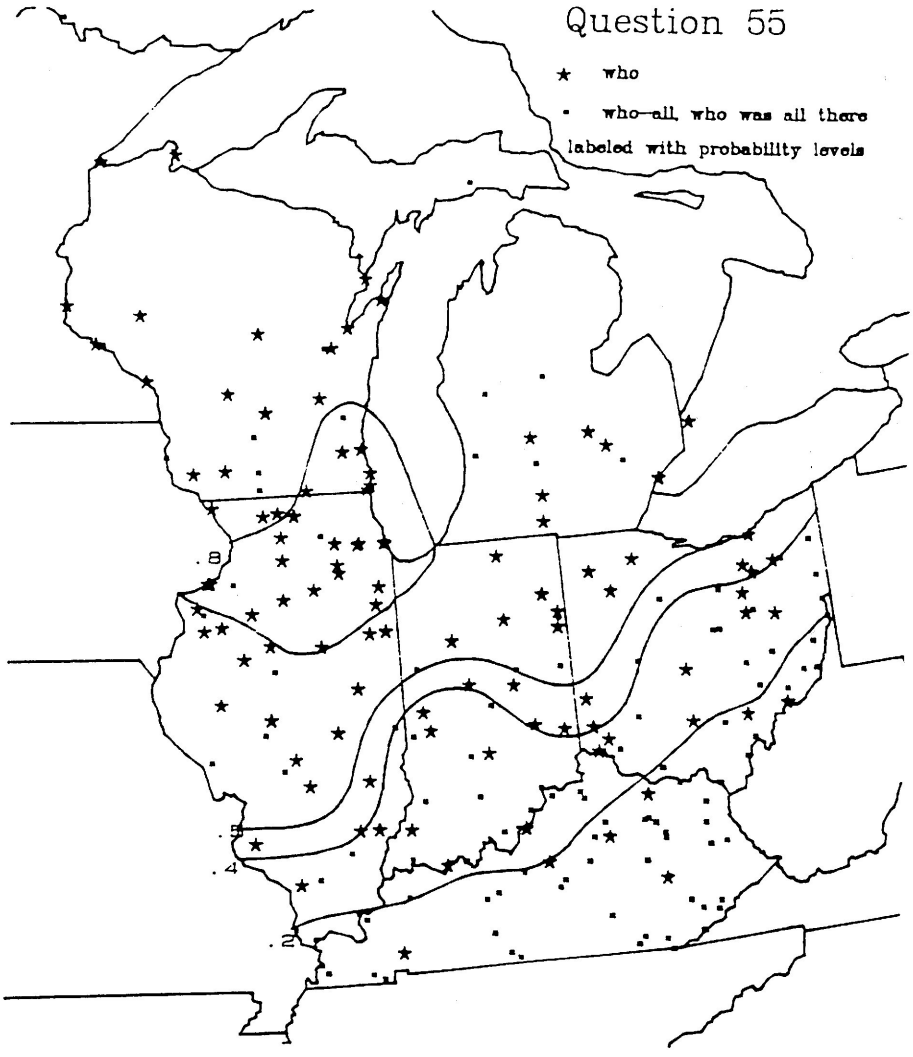


Karte 8: Strahlenkarte: Intervallalgorithmus: MINMWMAX

Figure 2-4 "Strahlenkarte" in Northern Italy by Goebl (1993)

remote ends of a country. These forms are interpreted as older forms in relic areas. As in **Figure 2-6**, in the fictitious case of the distribution of A / B / A, on an isolated island, isoglosses are drawn between word-forms A and B, and also between B and A. Isoglosses are drawn similarly when three different word-forms A / B / C are used in the same island. Thus neither the isogloss method nor the honeycomb method can show coincidence of phenomena in remote areas.

Here we should distinguish between dialect division and dialect classification. Dialect division can be executed by the ISOGLOSS METHOD without paying attention to coinci-



Map 5: Contours for Question 55: Who

Figure 2-5 A CONTOUR MAP of American Dialect by Linn and Regal (1993)

dence between remote areas. In dialect classification, linguistic similarities between dialects are taken into consideration. Of the methods which will be discussed later, both the IDENTITY METHOD and the MULTIVARIATE ANALYSIS take linguistic similarities into consideration. Thus they are the methods used in dialect classification.

In dialect classification, it is necessary to pay attention to linguistic features, in order to show similarities and differences of dialectal phenomena. If the use of a computer is considered the IDENTITY METHOD is generally simpler and easier to apply. Counting differences or similarities by computer is, on the whole, more efficient and more effective. Once research data is put into a computer, quantification is very easy. It is almost impossible to count a large amount of data of many answers from many localities by hand. This kind of processing is the field in which computers can accurately deal with a huge amount of data in a short time without mistakes.

2.2. IDENTITY METHOD

The IDENTITY METHOD is also called the similarity method or in Professor Vier-

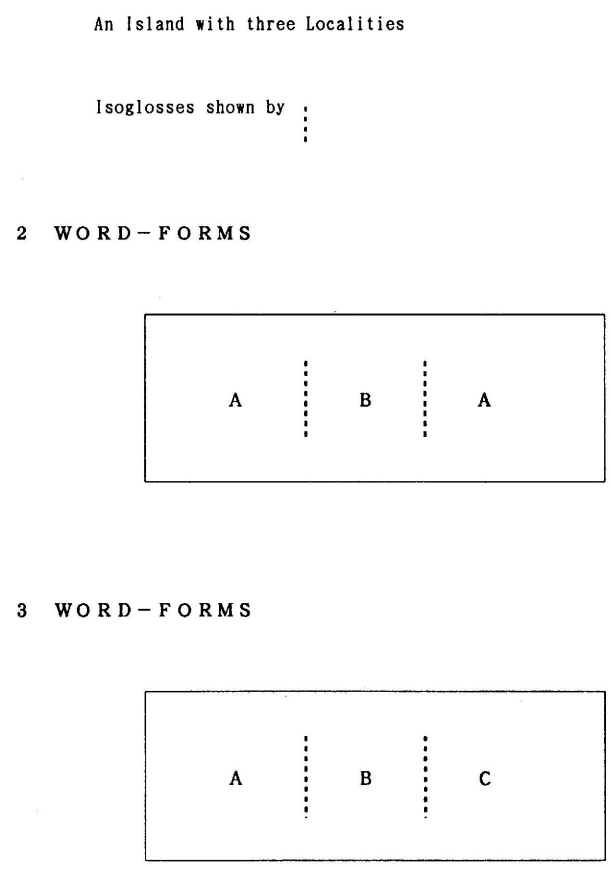


Figure 2-6 An application of the ISOGLOSS METHOD to an isolated Island dialect division and dialect classification.

eck's terms "Identity Test". In this method degrees of linguistic similarity are calculated and shown on a map. The simplest identity mapping of a geographical distribution is the encircling of a distribution area of a linguistic phenomenon. The cumulative identity method is useful for showing the general distribution pattern of an area as a whole. In the cumulative identity method identical phenomena in the localities in a number of maps are counted. As shown in the simplified data matrix of **Figure 2-7**, in making cumulative maps one counts and sums up the related items for each locality. Data is made for each locality but information for each word (item) is ignored.

The cumulative identity method without the use of a computer has a long tradition in Japan. My paper about Identity Method and Gravity Center Method using a computer was written in 1984. The identity method is used by Viereck (1992b), Hummel (1993) and Putschke (1993).

Some keys are necessary in order to group words for counting. Keys can be classified into two groups: one non-distributional and the other distributional.

Simplified Raw Data

	Localities	a	b	c
W	1	A	A	A
O	2	A	A	B
R	3	A	B	B
D	4	A	B	A
S	5	A	B	C

Cumulative Identity Value

Key = A

Localities	a	b	c
Identity Value	5	2	2
%	100%	40%	40%

Figure 2-7 The CUMULATIVE IDENTITY METHOD: Basic Procedure

1. Distributional keys. These are distribution patterns themselves found through the observation of maps. For example, western and eastern (or northern and southern) forms which may be found by observing each linguistic map. But here there is a problem of objectivity, of how to select the key distribution patterns or words without arbitrary selection. The most objective keys are individual localities.

The data presented here is based on a provisional grammatical survey of 153 localities organized by National Language Research Institute (Kokuritsu Kokugo Kenkyujo 1979) and was computerized by myself (Inoue 1984, 1990). This is an example of nominal (non-numerical) data. **Figure 2-8** is a map of identity method taking a southernmost locality (Kagoshima) as a key locality for comparison. The key locality, shown by a triangle, is naturally 100% identical to itself, but the neighboring localities show a very low identity value. This is quite different from the map in **Figure 2-9** which, starting from northern Japan (Aomori), shows high degrees of similarities with other localities. This suggests that linguistic diversity in southern Japan is greater than in other parts of Japan.

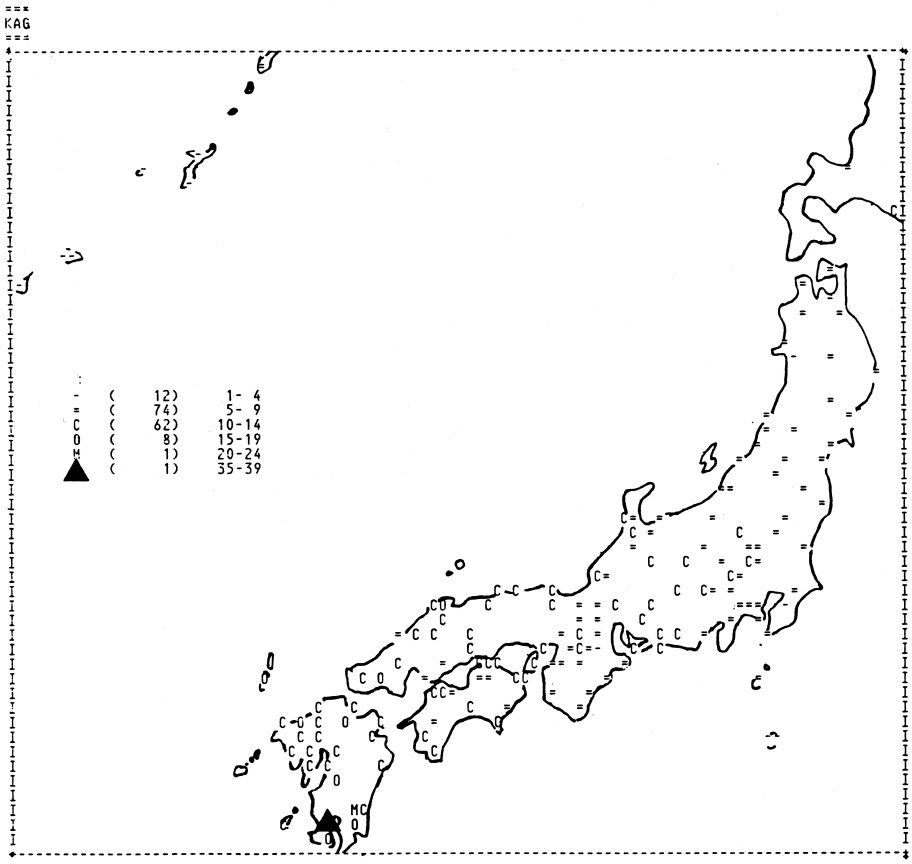


Figure 2-8 The CUMULATIVE IDENTITY METHOD: Grammatical data—— Kagoshima as a locality for comparison

2. Non-distributional keys. Standard language or historically older words can be selected as the keys to count identity.

Figure 2-10 shows the number of standard Japanese forms or rather common (colloquial) grammatical expressions used in each locality. This map shows the geographical background of standard Japanese. That is, that common Japanese grammatical forms are used mainly near Tokyo.

Figure 2-11 shows Fukushima's application of the cumulative identity method using a personal computer (Fukushima 1993). The key used here is the old pronunciation [a:] used in the northern area of western Japan. **Figure 2-12** is a map by Maekawa (1993) showing the percentage of appearance of an old phonological feature [kwa] in a neighboring area of western Japan.

2.3. CUMULATIVE IDENTITY MATRIX

The identity value between any of 153 localities of the grammatical data above can be calculated, and if the identity values of all the possible localities are calculated, a data

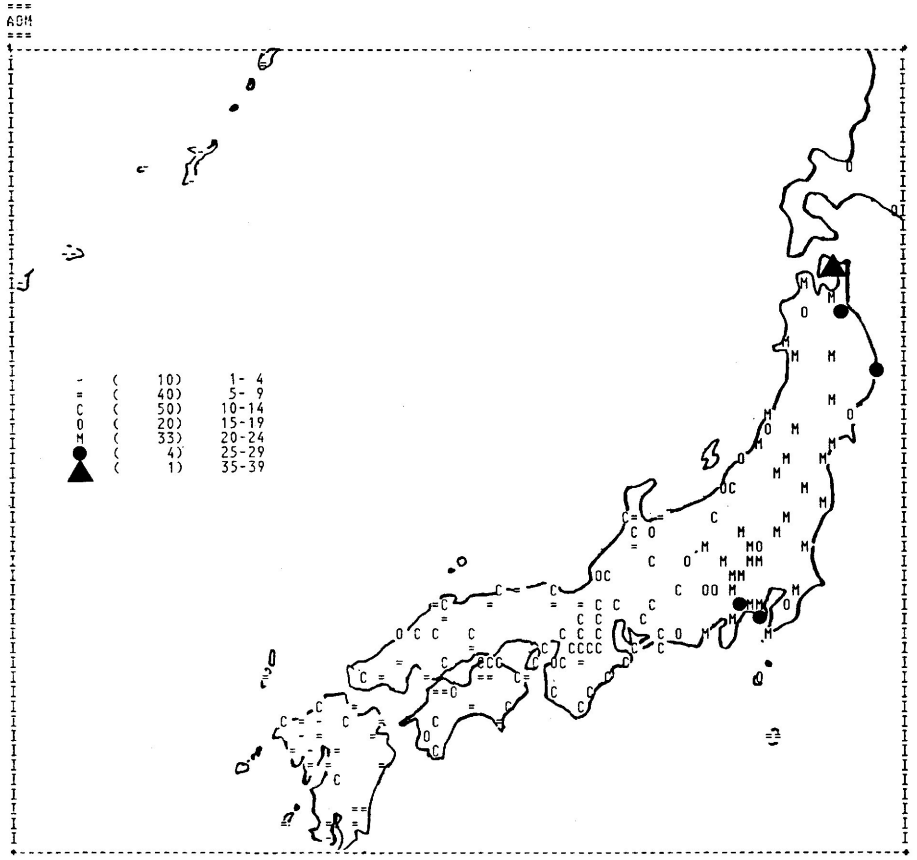


Figure 2-9 The CUMULATIVE IDENTITY METHOD: Grammatical data—— Aomori as a locality for comparison

matrix of all the localities can be produced. It is difficult to represent this on maps, because the same number of maps as the number of localities is necessary. However if we plan to apply multivariate analyses the data matrix is useful, because a matrix taking the form of numerical values is easier to process.

This cumulative identity method generally yields better results after the application of multivariate analysis, because after that all the forms can be classified without exception.

2.3.1. MAKING NUMERICAL DATA FROM NOMINAL DATA——IDENTITY MATRIX

The identity method thus supplies a numerical data matrix, which can be processed by multivariate analysis. Numerical, in the sense that data is represented by continuous numbers or numerical values.

The data can be converted into numerical form by counting identity values between every possible locality as pointed out above. There are many techniques to make numeri-

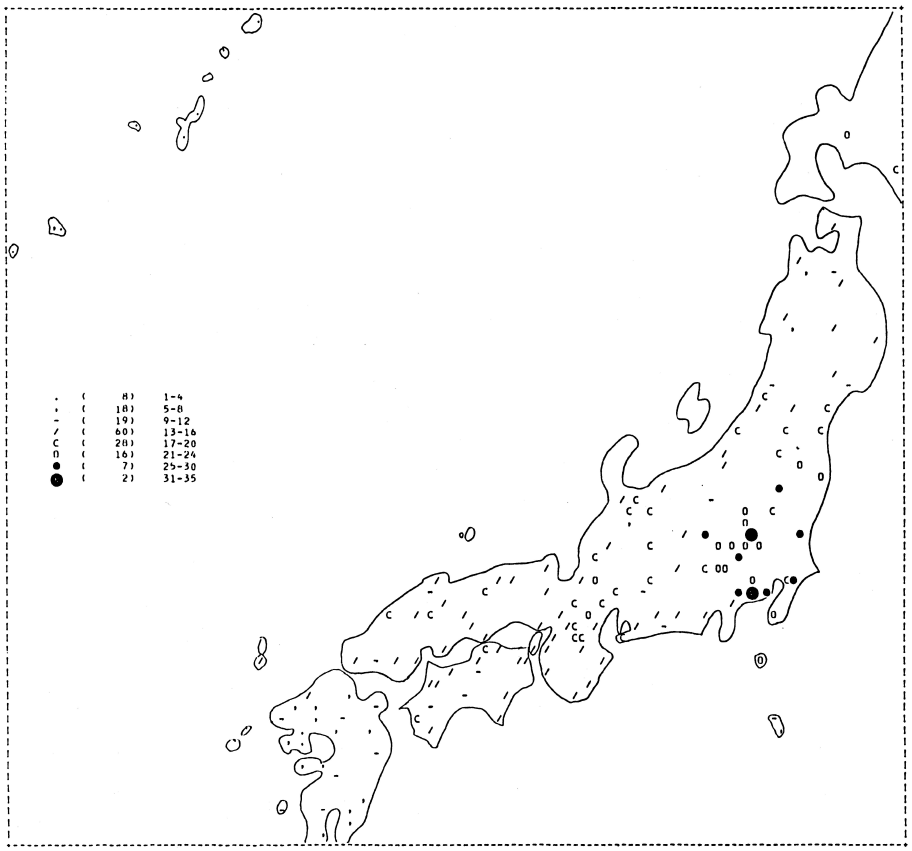


Figure 2-10 The CUMULATIVE IDENTITY METHOD: Grammatical data—— Standard Japanese forms for comparison

cal data from dialect distribution data.

The procedure of counting the same forms between many investigated localities has become possible because of computers. But the resultant data matrix is usually too large to be interpreted directly by human eyes. Some additional statistical procedure is necessary to simplify the multidimensional relations. Many scholars are now interested in applying multivariate analyses to the identity matrix of localities. Many attempts at statistical methods including cluster analysis are being made these days. Proceedings of the First International Congress in Bamberg (Viereck 1993) show some representative studies.

2.3.2. DIALEKTOMETRIE BY GOEBL

Figure 2-13 shows Professor Goebl's method (1993), counting the identity values between all the localities. The graphic technique of showing the continuum of similarities between the localities is well-designed. The continuum of similarity is shown by computational stereogram by means of the hypothetical height of mountains. Inclusion of a hypothetical point for standard language was also successful. But in this method it is almost impossible to show all the maps. Professor Sibata managed to show all the maps

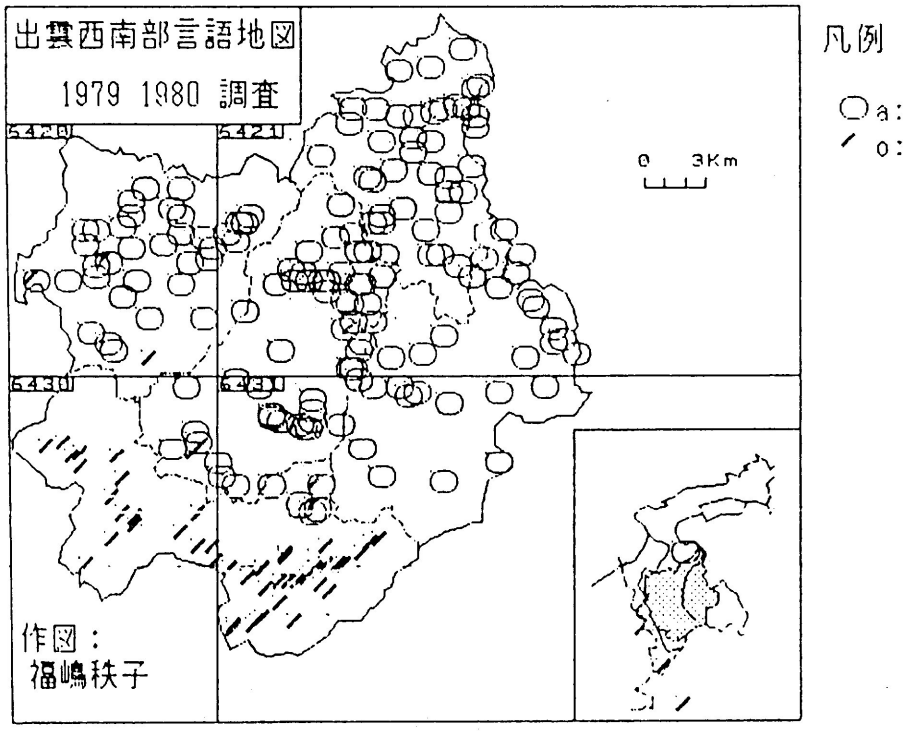
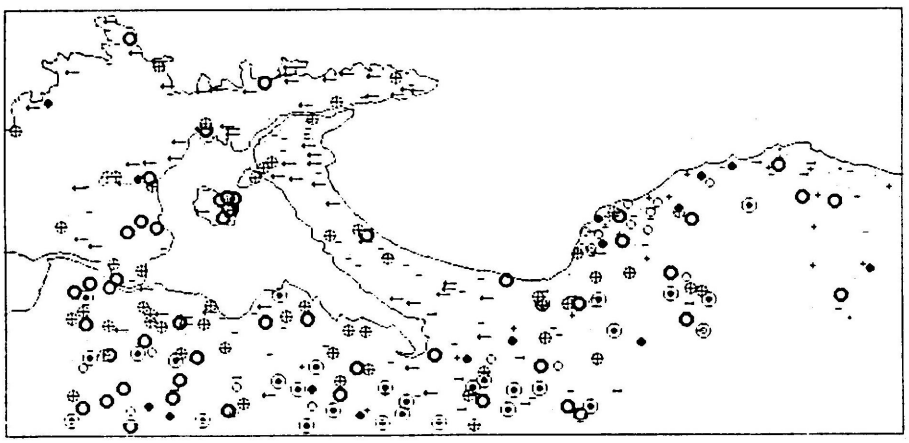


Figure 2-11 The CUMULATIVE IDENTITY METHOD by Fukushima (1991)
Old pronunciation [a:] and [o:] in a district of western Japan.

for the data of Amami-Oshima (Sibata 1984).

2.3.3. S & K NETWORK METHOD BY SIBATA & KUMAGAI

As for the technique of presenting identity matrices, the most effective practice in Japan is the "S & K Network Method" or the "N-K method" for dialect division developed by Sibata and Kumagai (1993). In this method lines connecting localities of a certain degree of identity (or similarity) are shown between all the localities. The method was applied to Amami-Oshima Island and later to central Japan in the Itoigawa area (Sibata & Inokuchi 1992). **Figure 2-14** shows a recent application to the data of the "Linguistic Atlas of Itoigawa". Application of this method was rather easy in spite of the large amount of data, because the data of the "Linguistic Atlas of Itoigawa" had been input into computer and the atlas is being published by computer. When data is already in computer it is easy to



記号割当 B:I W%.EGL　　F6:画面↓ F7:画面↑ F9:ﾒﾆｭｰ F:出力 C:ｺﾋﾟｰ

NO	語形		頻度	NO	語形		頻度
1	0 %		(35)	2	10 %	·	(4)
3	100 %	o	(50)	4	20 %	+	(21)
5	30 %	-	(53)	6	40 %	–	(16)
7	50 %	⋯	(63)	8	60 %	○	(22)
9	70 %	●	(16)	10	80 %	⊕	(55)
11	90 %	⊕	(29)	12			(0)
13			(0)	14			(0)
15			(0)	16			(0)
17			(0)	18			(0)
19			(0)	20			(0)
21			(0)	22			(0)
23			(0)	24			(0)
25			(0)	26			(0)
27			(0)	28			(0)
29			(0)	30			(0)

FIG.10 「合拗音の残存率」の言語地図

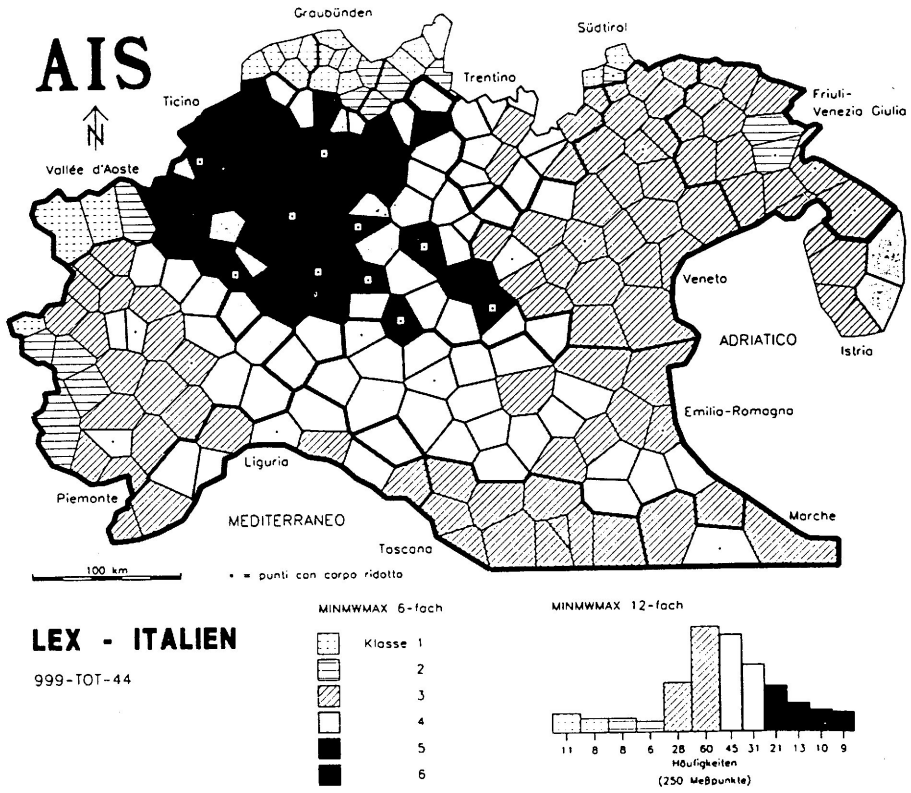
Figure 2-12 The CUMULATIVE IDENTITY METHOD by Maekawa (1988)
Percentage of the old pronunciation [kwa]
in a district of western Japan.

apply a newly developed method.

Both Professor Goebl's and Professor Sibata's data were also processed by other multivariate methods. We will be able to compare the results later.

As Figure 2-7 shows, the data matrix consists of two dimensions of linguistic items (dialectal forms, words or phenomena) and localities (informants). Theoretically an identity matrix can also be made in the dimension of linguistic items (words). In this case words will be classified according to patterns of geographical distribution.

Professor Cichocky (1993) reinforced Professor Viereck's suggestion that linguistic dimension should also be analyzed in dialect division in order to detect which linguistic features contribute to the divisions. Several multivariate analyses can play a decisive role for this purpose. The gravity center method, which will be discussed later, also has potential for this kind of analysis.



Karte 3: Beispiel für eine typisch lombardische Ähnlichkeitskarte in Choroplethentechnik

Prüfbezugspunkt: AIS-P. 44 (Mesocco, Graubünden, Schweiz).
Intervallalgorithmus: MINMWMAX (cf. dazu 3.2.).

Figure 2-13 The IDENTITY METHOD in Northern Italy by Goebl (1993)

2.4. AVERAGE IDENTITY MATRIX——THE KASAI DATA

There is another technique for cumulating answers. The Kasai data which will be utilized often hereafter is itself a result of the cumulative identity method (Inoue & Kasai 1989). As **Figure 2-15** shows, this time cumulation was executed in a different dimension of the data matrix. Localities were grouped by means of prefectures. Hisako Kasai prepared this data by counting and summing up standard Japanese forms for each prefecture for each word (item). The data matrix in Figure 2-15 shows the ratio of use of standard Japanese forms (here shown by "A") per prefecture. This kind of data may be called AVERAGE IDENTITY DATA. As the data was assessed and processed by prefecture, distribution within each prefecture cannot be retrieved and only a gross geographical pattern can be treated. But characteristics of individual words (items) can be distinguished. This is a departure from the Isogloss method which are concerned with individual localities. The Kasai data can also be processed by multivariate analysis as it is in the form of numerical values.

A part of the actual matrix of selected prefectures and selected words is shown in **Figure 2-16**. The data format shows percentage of standard Japanese forms as columns, calculated for the prefectures (rows). The entire matrix consists of 82 standard Japanese forms as cases, and 48 prefectures as variables (Inoue & Kasai 1989). The original data is based on 82 selected maps of the "Linguistic Atlas of Japan" (LAJ) including 2,400 localities.

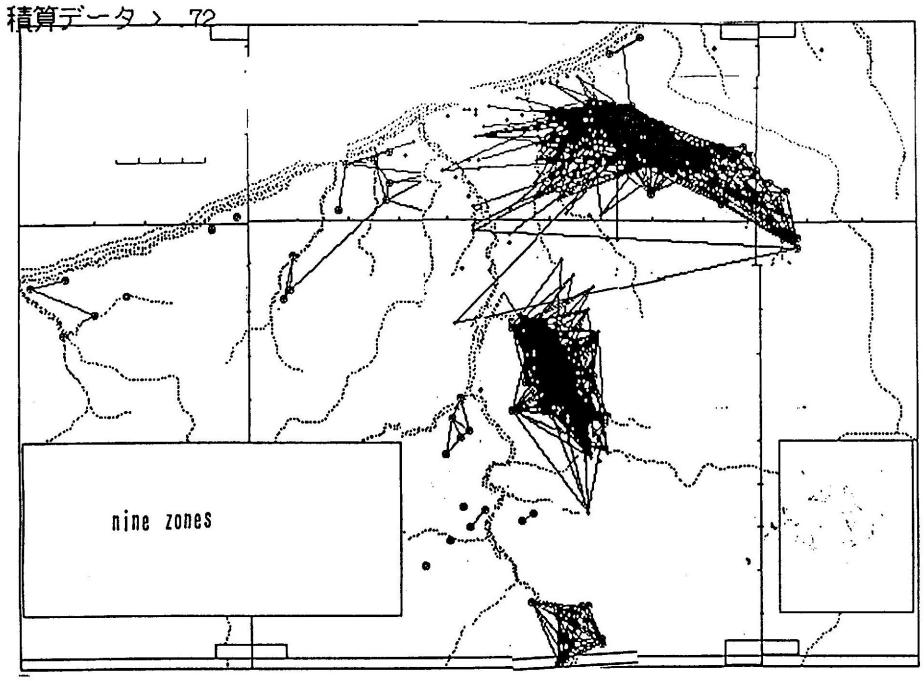


Figure 2-14 The S & K NETWORK METHOD by Sibata (1992)
Application to the "Linguistic Atlas of Itoigawa"

2.4.1. DIRECT RESULTS OF THE KASAI DATA: SIMPLE QUANTIFICATION

When the Kasai data was prepared by hand of Hisako Kasai, many maps were drawn up. This was an attempt to geographically present AVERAGE IDENTITY DATA. Firstly, distribution maps of the percentage of usage of standard Japanese forms were made. **Figure 2-17** is a sample map of the percentage of standard forms for the word meaning "icicle". From the data matrix for 82 words, 82 maps of this kind were made respectively.

As shown in **Figure 2-18** the average percentage of standard Japanese forms as a whole was calculated and shown for each prefecture. This is one more step of averaging from average identity maps. This map shows that standard forms are mainly used near the capital of Tokyo, and that there is a cleavage of standardization between Eastern and

A = STANDARD JAPANESE

		X PREFECTURE				Y PREFECTURE				
	Localities	a	b	c	d	e	f	g	h	i . . .
W	1	A	A	A	A	A	A	A	A	
		100%				100%				
O	2	A	A	A	A	B	B	B	B	
		100%				0%				
R	3	A	B	B	B	B	B	A	A	
		25%				50%				
D	4	A	A	B	B	A	A	A	A	
		50%				100%				
S	5	A	B	B	B	C	C	C	C	
		25%				0%				

Figure 2-15 AVERAGE IDENTITY DATA——Basic procedure of the Kasai Data

Western Japan. This distribution pattern is similar to the result of the cumulative map of standard grammatical forms shown in Figure 2-10.

These maps were drawn up by hand. The Kasai data was later put into computer by Inoue and a variety of statistical techniques have since been applied to the data (Inoue & Kasai 1989).

3. MULTIVARIATE ANALYSIS——INTRODUCTION

The two quantitative methods discussed above are arithmetic and may be applied by hand if one has a calculator and enough time. But the multivariate or multidimensional analyses which will be discussed in this section are almost impossible to perform by hand even if one has a calculator and enough time. Sometimes a large mainframe computer is necessary to perform the calculation of multivariate analysis of a large number of words. The multivariate (or multidimensional) approach is thus a typical field of computational dialectology. It is a field which has been enabled by the recent developments in computers, that is hardware. The handling of software or programs has also become much easier for scholars of the humanities because of the increased availability of application programs or package programs.

Standard Japanese Forms	PREFECTURES										
	Hokkaido	Aomori	....	Tokyo	....	Kyoto	Hyogo	....	Okinawa
Mabushii	28.5	6.8	....	88.9	....	11.1	11.3	....	0.0
Koge-kusai	95.2	58.1	....	100.0	....	97.2	87.3	....	0.0
Nasu	15.4	0.0	....	100.0	....	18.8	18.2	....	0.0
Tsuyu	39.8	0.0	....	22.2	....	100.0	98.6	....	0.0
:	:	:		:		:	:		:
:	:	:		:		:	:		:
:	:	:		:		:	:		:

Figure 2-16 A part of the Matrix of the Kasai Data
Selected prefectures and selected words

Here I will describe some methods for analysis of numerical data and next methods for non-numerical data.

3.0. CORRELATION COEFFICIENT OF NUMERICAL DATA: THE KASAI DATA

When the data is numerical, many statistical methods are applicable, especially various multivariate analyses which have been widely applied in other research fields. When dialectological survey data has been converted into numerical form, many possibilities arise. For example, the following methods have recently been effectively applied to numerical data: Multiple Regression, Factor analysis (Inoue & Kasai 1989), PCA=

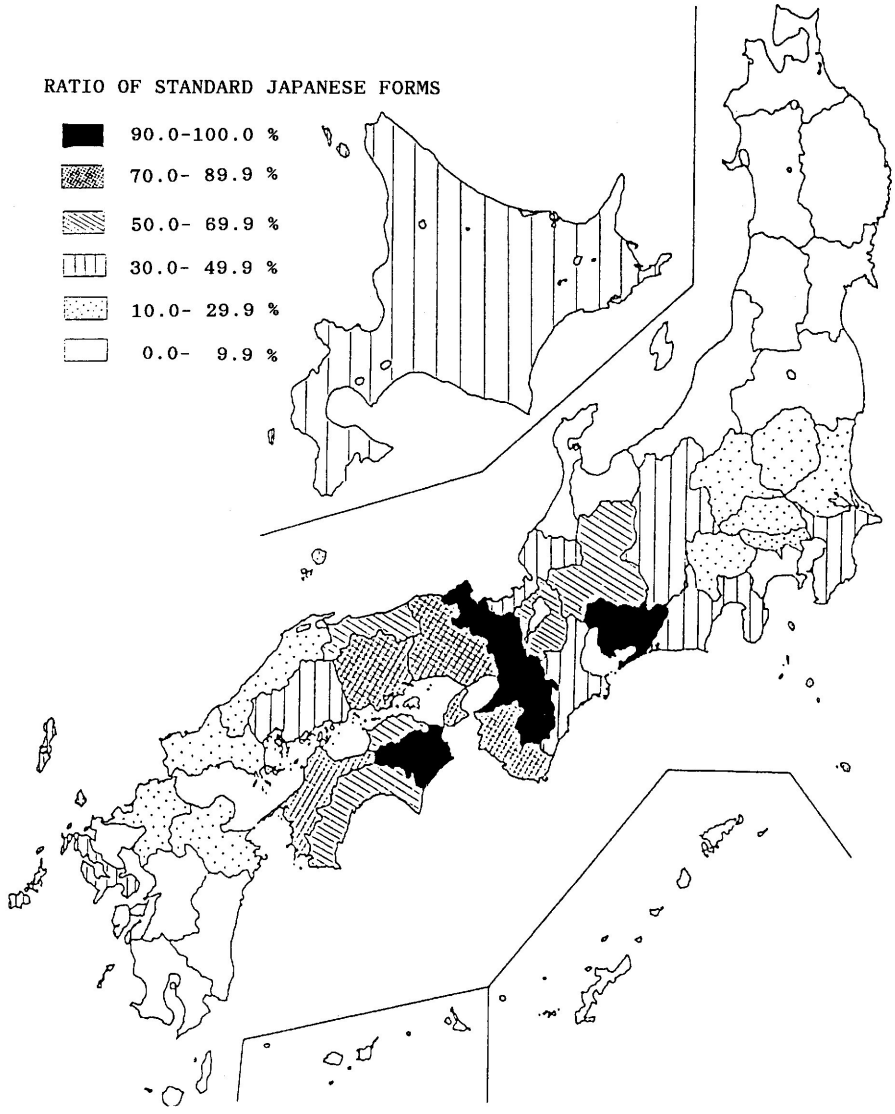


Figure 2-17 Sample Map of the Kasai Data——"Icicle"
Percentage of standard expression [tsurara]

Principal Component Analysis (Ogura 1990), MDS=MultiDimensional Scaling (Embleton 1987, Hoppenbrouwers 1993, Ainsaar & Ross 1993), Cluster Analysis (Viereck 1992, Klemola 1990, Funk 1992) and Dual Scaling (Cichocky 1993).

One problem in computational dialectology is that ordinary dialectological data is not in the form of numerical values but in the form of non-numerical, nominal, or the categorical answers of informants. When the data is converted into numerical form, the possibility of application of statistical methods increases. A good example of Japanese dialect data is the Kasai data.

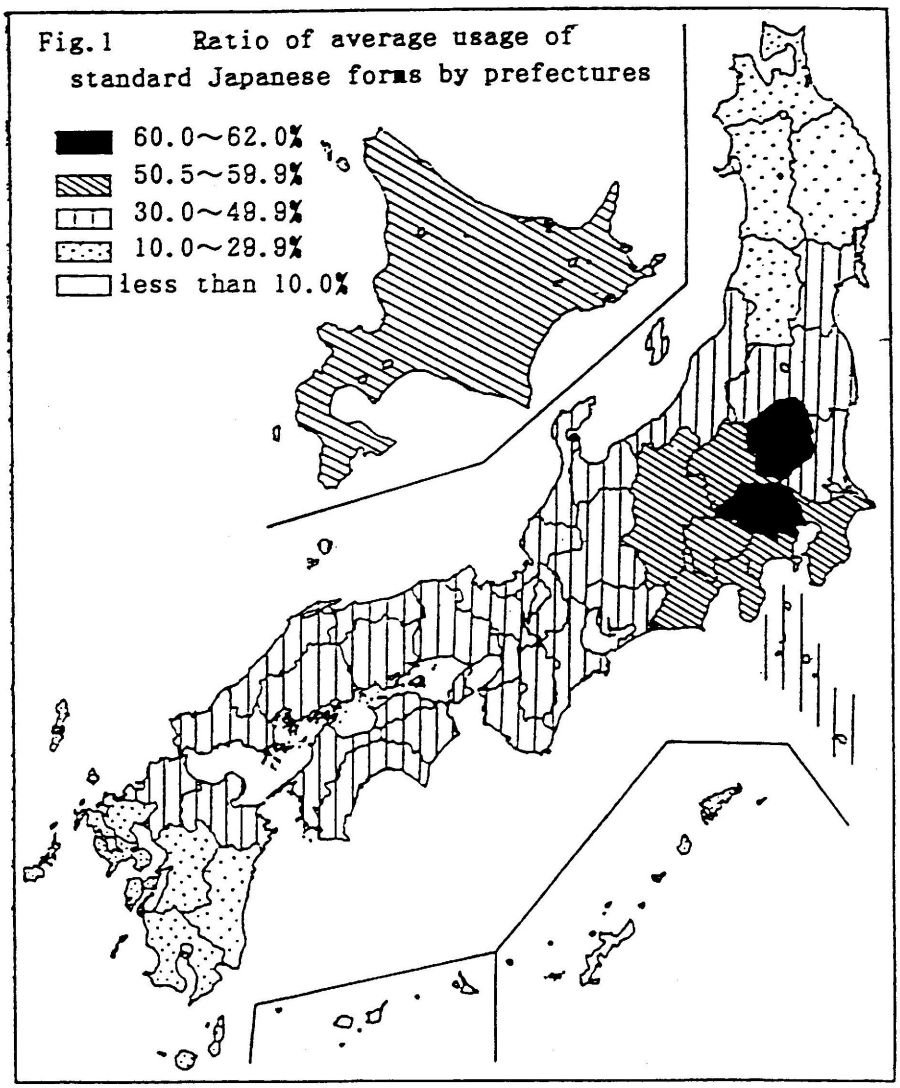


Figure 2-18 Average Percentage of Standard Japanese Forms of 82 words of the Kasai Data

This kind of matrix is too large for a human to grasp all in one. As shown by **Figure 3-1**, similarities between prefectures in the form of percentage of standard Japanese words can be shown more compactly by correlation coefficients which are calculated on the basis of the data above. The similarities between any prefecture and the rest of the prefectures can be shown in this table. Values in bold letters are higher correlation coefficients of more than 0.5, showing a close relationship between the prefectures. Several groups according to administrative districts can be identified.

But the pattern is still too complicated, and can be further simplified by computer analysis if multivariate analysis is applied. Machine readable data can easily be processed by multivariate analyses. The great effort of data input can be justified if the data are processed by multivariate analyses. The results are simpler and easier to interpret when one becomes accustomed to the methods.

3.1. MULTIVARIATE ANALYSIS OF NUMERICAL DATA——CLUSTER ANALYSIS OF THE KASAI DATA

In the beginning of application of multivariate analyses CLUSTER ANALYSIS will be discussed. The basic idea of cluster analysis is similar to and most appropriate for dialect

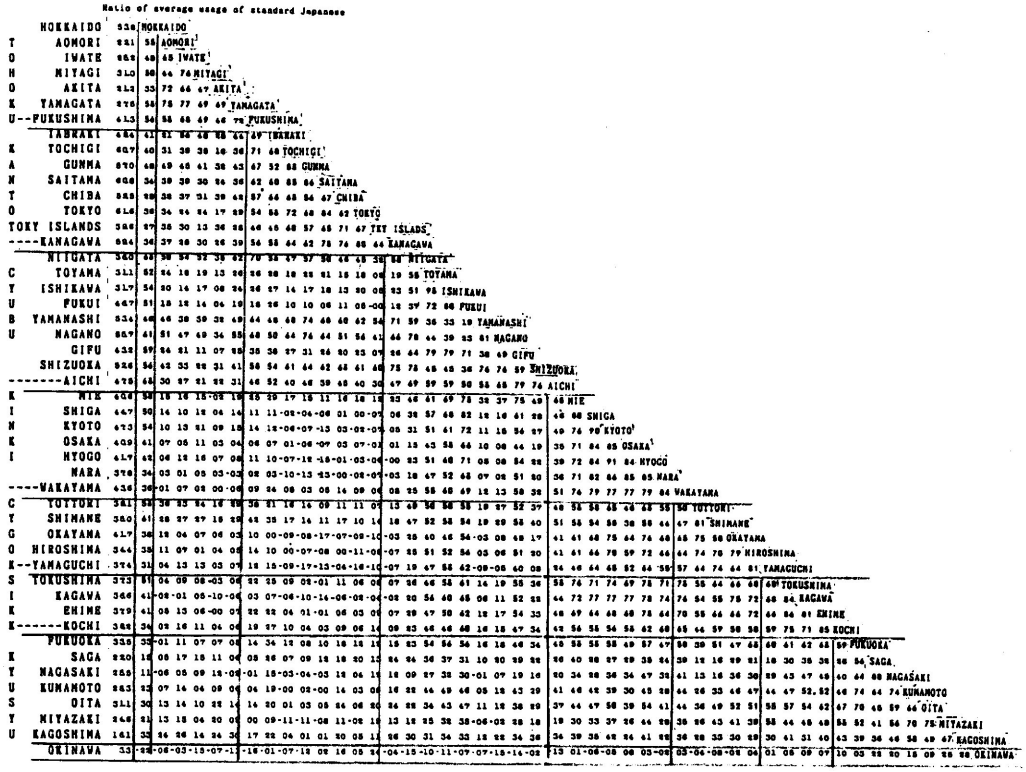


Figure 3-1 CORRELATION COEFFICIENTS of the Kasai Data among 48 Prefectures

division. It has been applied by Goebl (1993), Sibata & Kumagai (1993) and Shaw (1974) to their identity matrices, producing plausible results.

The merit of cluster analysis is that it offers an objective method of dialect division. Divisions can be based on many phenomena, thus avoiding arbitrary and subjective factors in analysis.

Cluster analysis is more appropriately applied to numerical data than to nominal data. Identity matrices such as those of the Kasai data may be finely processed by cluster analysis as shown in **Figure 3-2** (Inoue & Kasai 1989), and classification by CLUSTER ANALYSIS will be used for reference in presenting the results throughout this paper.

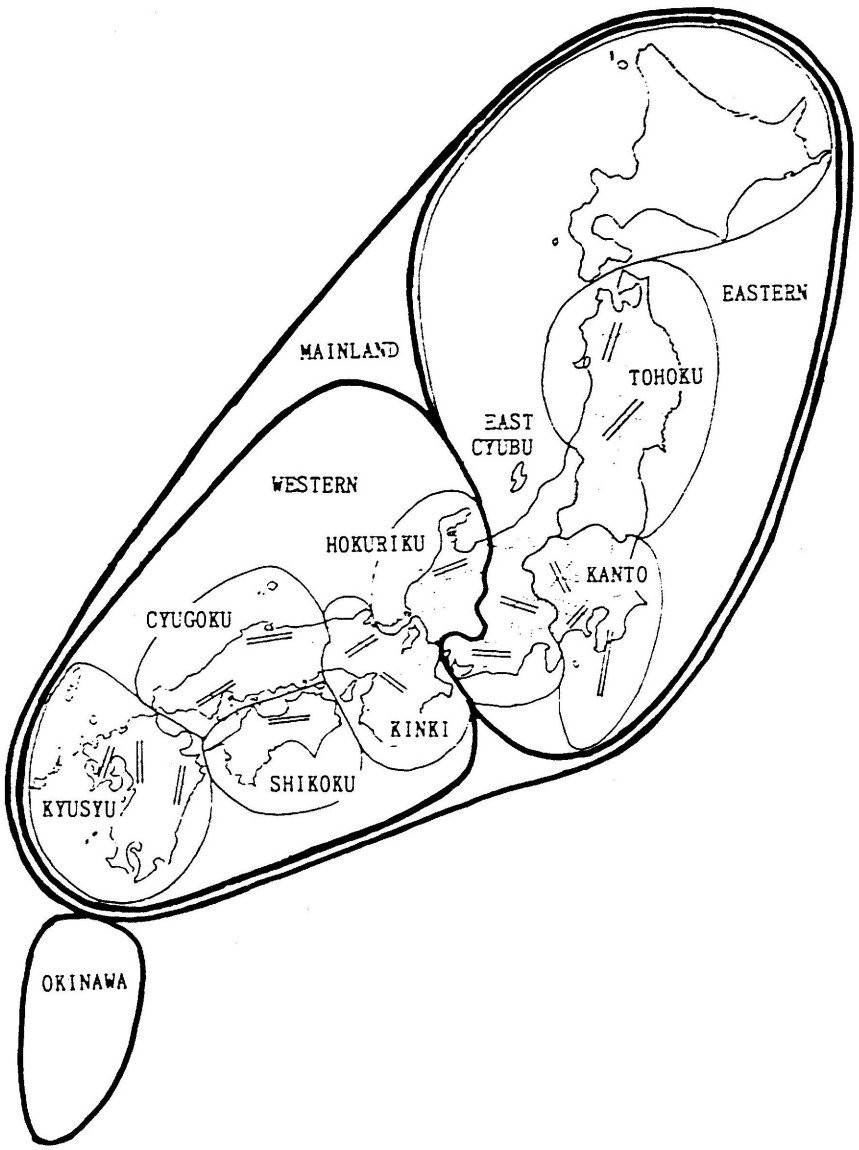


Figure 3-2 CLUSTER ANALYSIS of the Kasai Data for 48 prefectures

The procedure of calculation of cluster analysis will be explained on the basis of the Kasai data. Basically the answers for all the prefectures are compared with each other and the most similar pair of prefectures is combined as one cluster. These two are hypothetically treated as one prefecture. Next, all the data with one less prefecture is compared again, and the most similar pair is again combined as the next cluster. This process continues until the last two hypothetical prefectures are combined as one cluster. There are many techniques for calculating similarities and for making a new hypothetical prefecture.

In the case of the Kasai data, cluster analysis can be applied not only in the dimension of prefectures but also in the dimension of items or words. These are the Q technique and the R technique. Both techniques were applied.

3.1.1. CLASSIFICATION OF PREFECTURES

The map in Figure 3-2 shows the result of cluster analysis of prefectures. Many techniques were applied and the results were all a little different. The simplest (average linkage) technique was selected to be presented here. The hierarchy of clusters is shown by the circles of different size and the nearest pairs are joined by two lines.

The resultant dialect division by cluster analysis is very plausible. On the whole the results coincide with the previous attempts at dialect classification, that is, the separation of Ryukyu (Okinawa) dialect from the mainland dialect, and the division of the mainland dialect into the western and eastern dialects. But the location of the border of east and west is different from earlier studies. Here the mountainous area in central Japan is the dividing line between east and west. As for the smaller divisions, it shows coincidence with the usual, educational divisions of Districts in Japan. Although the data itself was independently collected, the patterns of answering (use of standard Japanese forms) is similar for each District.

The map of Figure 3-2 thus provides reasonable results for dialect division, based on the ratio of usage of standard Japanese forms.

3.1.2. CLASSIFICATION OF WORDS

Cluster analysis can also be applied in the other dimension of the matrix, that is words (or items). Standard Japanese words similarly used in prefectures were grouped together by this procedure.

Figure 3-3, the so-called DENDROGRAM, shows that all the words are divided into two main clusters. Comparison with individual dialect maps shows that they are words with eastern and western distribution. The eastern sub-clusters are further divided into 3 sub-clusters A, B, C, the western sub-clusters into 6 sub-clusters D, E; F, G; H, I. As A subcluster is larger than the others, sub-clusters of A were named from A1 to A4 subclusters. Comparison of the clusters with each linguistic map showed that the classification by cluster analysis represents geographical distribution quite well. As will be shown later, the result of cluster analysis will be utilized hereafter, by labeling each word with the names of the sub-clusters.

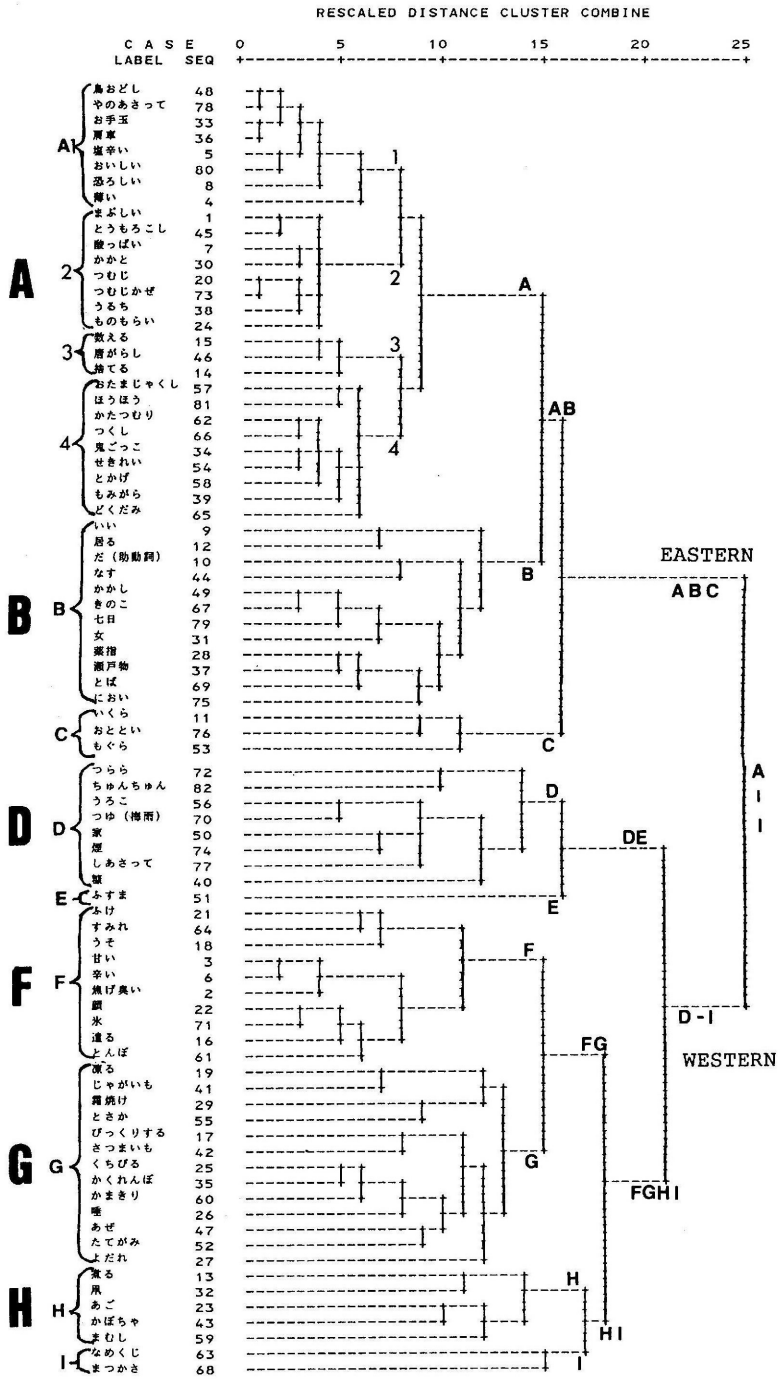


Figure 3-3 CLUSTER ANALYSIS of the kasai Data—— Dendrogram of 92 lexical items

3.1.3. CLUSTER ANALYSIS OF NOMINAL DATA

Cluster analysis can be applied to non-numerical, nominal or categorical data, that is data without continuous numerical values. One kind of cluster analysis, "numerical taxonomy" used by Linn and Regal (1985), is a case in which responses are in principle represented by 1 or 0 corresponding to use or non-use of respective word-forms. There are several more studies in the same vein (Kelle 1986). I have also applied cluster analysis to various nominal data, but the results were not plausible (Inoue forthcoming). It is perhaps because there are only two values (1 or 0). Some kind of transformation of the data seems necessary before applying cluster analysis to nominal data.

3.2. MULTIVARIATE ANALYSIS OF NUMERICAL DATA——FACTOR ANALYSIS OF THE KASAI DATA

Another effective multivariate approach for numerical data is FACTOR ANALYSIS (see Figure 3-4). In factor analysis a whole pattern is explained as the conglomeration of many factors. It is literally a typical multi-dimensional analysis. Factor analysis is basically a method for analysis or characterization and not for division. This method was developed initially in psychology. Human psychological phenomena was thought to be composed of many components or factors. This idea has similarity with "Wellentheorie" of Johannes Schmidt and is appropriate for complicated dialect distribution with many waves of diffusion. Factor analysis generally gives natural and plausible results. Principal Component Analysis (PCA) is more appropriate if one wants to further simplify the geographical distribution. But in our analysis of the Kasai data identification of various distributional factors was more important.

The merit of factor analysis is that it is possible to extend the statistical characteristics of the dialects when other information is added to the data. Now studies of the Kasai data are in progress concerning historical factors on the basis of the earliest incidence of words in historical documents, relations with lexicological characteristics, communication factors on the basis of telephone traffic matrix and correlation with geographical distances (Inoue 1990). The demerit of factor analysis is that dialectological raw data which is by nature non-numerical should be converted into numerical form.

3.2.1. CLASSIFICATION OF PREFECTURES BY THE KASAI DATA

Factor analysis was applied to the Kasai data. Classification of prefectures can be shown by values of FACTOR LOAD. The map of **Figure 3-4** shows factor loads of the first to fourth factors for 48 prefectures. Prefectures which showed a factor load of more than 0.8 and more than 0.5 for the first and second factors, and prefectures with a factor load of more than 0.5 for the third and fourth factors are shaded differently on the map.

The first factor can be called the western (or Kansai) factor, because the rate is high in the western districts around Kyoto and Osaka. This shows that there is a large communality between western prefectures as to the use of the standard Japanese forms. Another study showed that this factor corresponds with standard forms which have spread from the old cultural center of Kyoto in the middle ages (Inoue 1990).

The second factor can be called the eastern (Kanto) factor, because it reflects the large correlation coefficients in eastern Japan around Tokyo. This factor corresponds with the diffusion of standard Japanese forms from the new cultural center, Tokyo, in modern times.

Comparison with the results of cluster analysis in Figure 3-2 shows that cluster analysis gave a clear-cut result. Factor analysis in contrast shows more subtle and complicated relations between the prefectures. This is especially so when we take into consideration the third and fourth factors, which correspond to northern (Tohoku) and southern (Kyushu) factors.

3.2.2. CLASSIFICATION OF WORDS BY THE KASAI DATA

Using the other dimension of the data matrix, classification of words can also be attempted by means of factor analysis. The values of FACTOR SCORES show the characteristics of the words investigated. **Figure 3-5** is a graph showing the combination

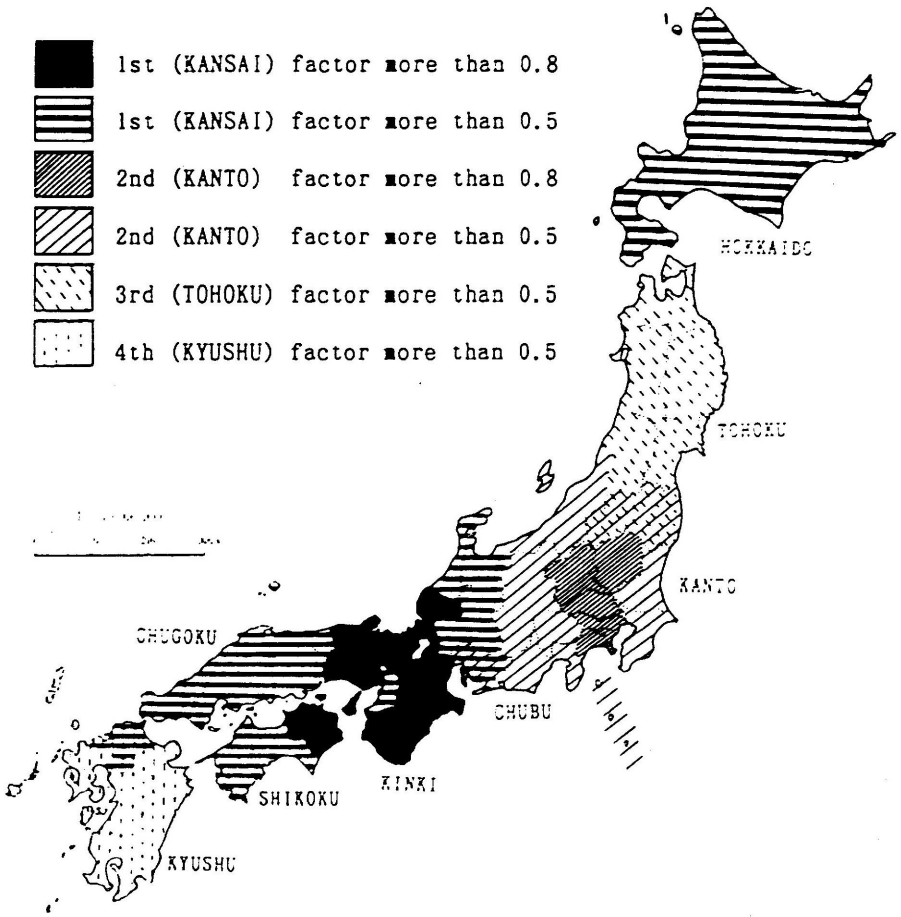


Figure 3-4 FACTOR LOAD of the First to Fourth Factors—— the Kasai Data for 48 prefectures.

of factor scores of the first and second factors.

As can be guessed from the characteristics of the two factors representing west and east, words in the four quadrants can be characterized as follows:

upper right means both eastern and western factors are strong, thus words distributing countrywide;

upper left means eastern factor is strong, thus words distributing mainly in the east;

lower right means western factor is strong, thus words distributing mainly in the west;

lower left means neither eastern nor western factors are strong, thus words without dialectal basis, mainly literary expressions.

These characterizations were mostly appropriate for the standard Japanese words when compared with individual distribution maps. When words are marked by the classification

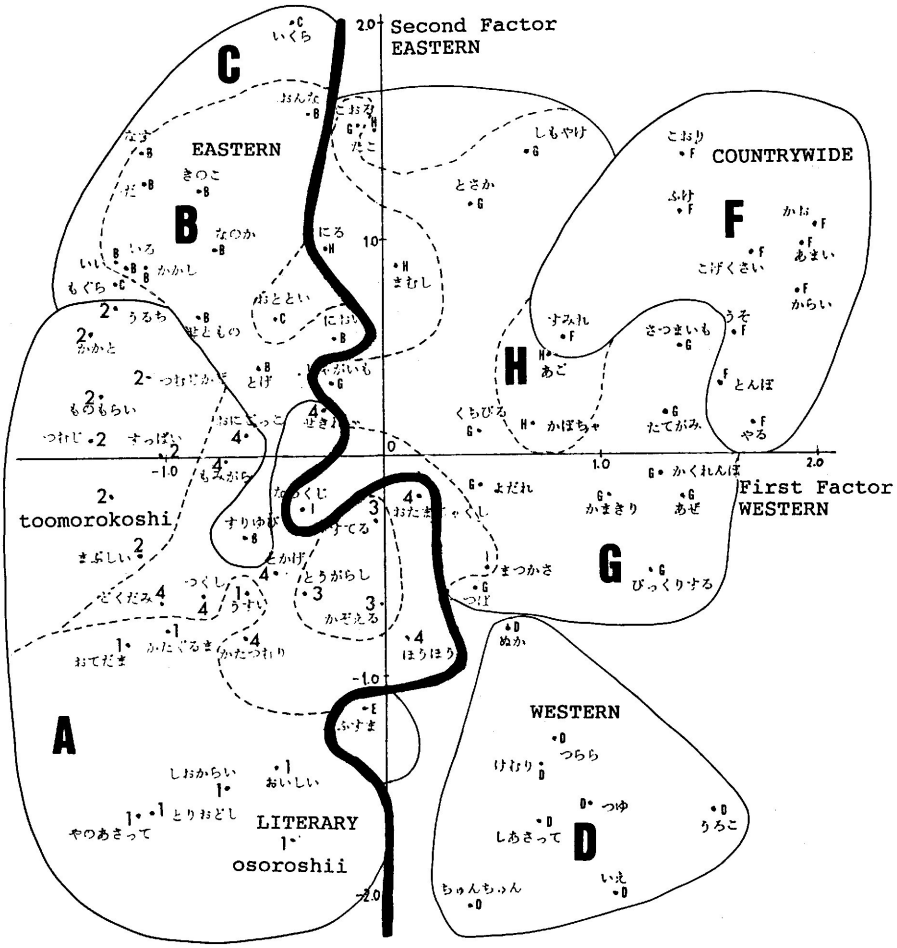


Figure 3-5 FACTOR SCORE of the First and Second Factors—— the Kasai Data for 92 lexical items

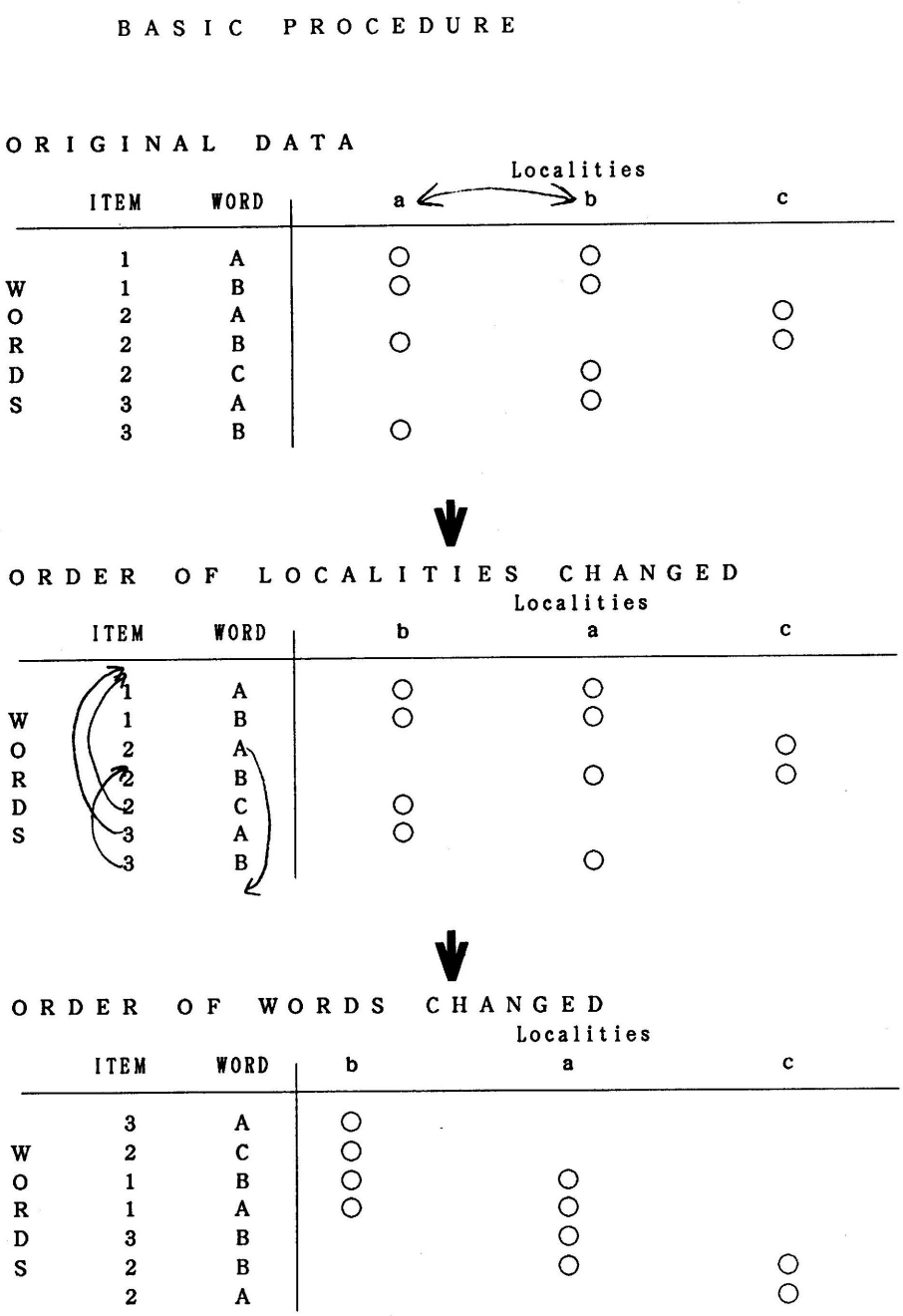


Figure 3-6 Basic procedure of HAYASHI 3
The order of both dimensions of the raw data are changed

of cluster analysis, as is shown in Figure 3-5, they basically coincide.

Cluster analysis does not show the continuum of dialectal distribution. Its results are sometimes too simple and much too clear-cut. Thus, factor analysis shows the basic trends of dialectal diffusion in Japan more faithfully.

3.3. MULTIVARIATE ANALYSIS OF NOMINAL DATA——HAYASHI'S QUANTIFICATIONAL THEORY TYPE 3

So far the statistical methods applied to the data had to be processed before analyzing them by computer because numerical data is usually easier to process. Now I will introduce a method by which raw data of ordinary dialectological surveys can be directly processed by computer.

For nominal (non-numerical, categorical) data which are acquired in linguistic research, "Hayashi's Quantificational Theory" which was developed by a Japanese statistician C. Hayashi, works effectively. "Hayashi's Quantificational Method Type 3" (hereafter referred to as "HAYASHI 3") is a multivariate analysis which was developed for non-numerical (nominal, categorical) data. It gives similar results to factor analysis. The method is also called "quantification for grouping", or a "method for gathering similarities" (Hayashi 1954, Inoue 1986). This method is more suited for "dialect classification" than for "dialect division".

Figure 3-6 shows the basic procedure of Hayashi 3. The order of both dimensions of raw data (localities and words) are interchanged so that the correlation coefficients in the

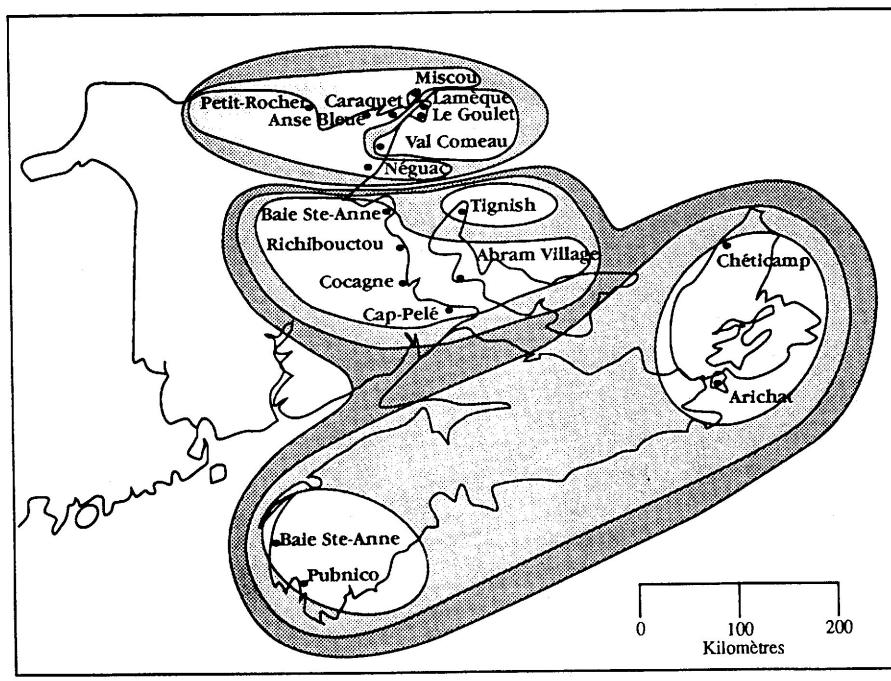


Figure 3-7 CORRESPONDENCE ANALYSIS of Canadian French by Cichocky (1992)

diagonal direction become the largest.

This method is analogous to the correspondence analysis developed by Benzécri in France (Benzécri 1981). Nishisato's method or Correspondence Analysis (in **Figure 3-7**) which Professor Cichocky and others applied to their Canadian French data is also similar in basic ideas. They used numerical values based on the incidence of many linguistic features. It is a method by which the two dimensions of linguistic features and localities (informants) can be processed at the same time.

"Hayashi 3" has been widely utilized in the social sciences and linguistic research in Japan. This method was first applied to sociolinguistic data and later to several geographical distribution data (including glottograms or age-area graphs), and clear distribution patterns were derived.

There are many examples of practical application. Here, a typical example of the application to Grammatical distribution data will be shown.

3.3.1. CLASSIFICATION OF WORDS BY GRAMMATICAL DATA

The dimension of linguistic phenomena (words) will be shown first. **Figure 3-8** shows some common Japanese grammatical forms. Because of the large size of the data, only

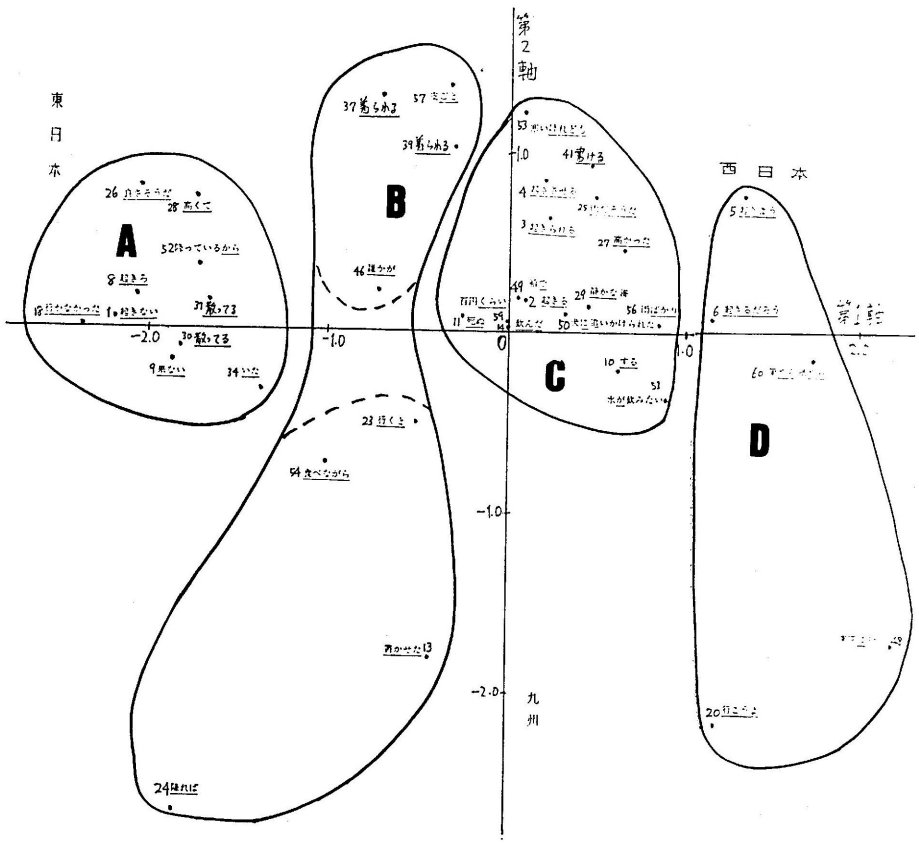


Figure 3-8 Results of HAYASHI 3——the Distribution of Grammatical Data

standard (or rather, common) Japanese forms were selected for analysis. As a result of Hayashi 3, standard Japanese forms have been divided into four main groups according to the value of the first (horizontal) axis. Comparison with original maps and the gravity center map in Figure 4-6 showed that four groups correlate with geographical distribution. All the forms have been geographically classified without exception as a result of Hayashi 3.

3.3.2. CLASSIFICATION OF PREFECTURES BY GRAMMATICAL DATA

The dimension of localities will be shown next in **Figure 3-9**. By calculating average values of the first and second axes for individual localities, dialect classification also becomes possible as a result of Hayashi 3. The values of the first axis are shown in the map of Japan in Figure 3-9. Preliminary application of Hayashi 3 to the whole data including dialectal forms showed that Japanese dialects are first classified into Ryukyu and mainland dialects. By the next calculation without the Ryukyu data, the mainland dialects have been classified into east and west by the values of the first and second axes. Next Kyushu (south) and the small island Hachijo were given independent positions. This division basically coincides with the lexical one discussed above in Figure 3-4. It is similar to a division advocated nearly a century ago by Prof. Misao Tojo, a pioneer of Japanese dialectology. However he later abandoned this division.

Comparison with individual maps showed clear geographical tendencies. Cumulative identity maps for the four respective groups of words were made. Figure 4-6 (which will be shown later) is a revised map. It shows that the groups are surely geographically

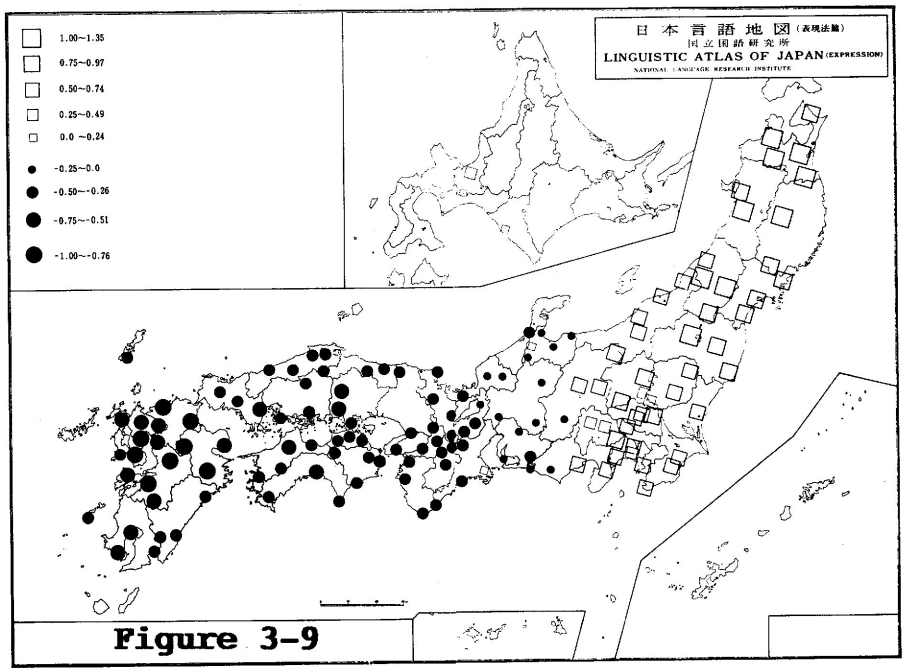


Figure 3-9 Results of HAYASHI 3——Map of Japan Grammatical Distribution Data

governed. Hayashi 3 was successful in bringing general distribution patterns into light. The difference between these cumulative maps and the traditional cumulative method is that all the words are classified in groups without exception. Accordingly, we can avoid arbitrariness.

3.3.3. MERITS AND DEMERITS OF HAYASHI 3

The greatest merit of Hayashi 3 is that the whole linguistic data can be processed directly by computer. No process of transforming non-numerical data into numerical values by counting identity or calculating averages is necessary. It enables us to directly find the patterns which are inherent or immanent in the data. Extralinguistic factors such as age, sex and location are not necessary at the time of calculation. The demerits are that a large computer memory is necessary, the data in alphabetical notation must also be converted into categorical numbers. Another demerit is that a category with only a few incidence sometimes skews the whole pattern.

3.4. TEMPORARY CONCLUSION

3.4.1. EVALUATION OF MULTIVARIATE ANALYSES

There are many more methods of multivariate analyses. They have been utilized in recent sociolinguistic research, including Principal Component Analysis (PCA) (Horvath 1985) in Australia, VARBURAL (Sankoff & Labov 1979) in America and Ogino's Quantification Theory (Ogino 1986) in Japan, to name only a few. All of them can also be applied to geographical distribution data in dialectology. The only difference is that in dialect geography informants are first and foremost characterized by localities, not by social class, age or sex. It is a great pity that I must omit the sociolinguistic studies of dialects because of my time limit, though I know that a large sociolinguistic survey has been conducted in Hungary (Kontra 1992).

As a result of the application of multivariate analyses, dialect classification has become possible while taking many words into consideration at one time. At the same time simplification of large amounts of data has become possible and the inner structure of overall patterns can be easily seen. I would like to advocate here that Hayashi 3 which can directly process the acquired dialectal data, is the best method for dialectology.

One practical problem of multivariate analyses is that it sometimes requires the large memory of a mainframe computer. But once the overall pattern has been grasped, simple mathematical calculations can show equivalent results if we pay attention to main characteristics. If simpler calculation affords similar results, this method will be more useful to many dialectologists.

3.4.2. FROM MULTIVARIATE ANALYSIS TO THE GRAVITY CENTER METHOD

As shown above in Figure 3-6 most of the multivariate analyses can be applied to the two dimensions of a data matrix: that is, to the dimension of localities (informants) and the dimension of words (linguistic forms). Thus multivariate analyses can be utilized both for dialect division and for classification or characterization of dialectal words. As was

discussed by Professor Cichocky and Professor Viereck we should elucidate at the same time what items or phenomena contributed to the given dialect division or distribution pattern. Hayashi 3 seems to be the best method for this purpose for the time being because it can perform both tasks at the same time.

For the former dimension of localities or informants, the identity method and isogloss method can contribute much with simpler calculation. As for the latter dimension of classification of words (or linguistic forms) some other statistical method should be thought of. A good candidate is the gravity center method. This method has another merit. Multivariate analyses including Hayashi's Quantificational Theory do not take geographical factors into calculation. But the gravity center method does.

This is the first half of the paper read for a lecture on the occasion of the First International Congress of Dialectologists and Geolinguists held in Budapest from 26th to 29th April, 1993. The Proceedings of the Congress does not seem to appear for a time being because of economical difficulties. The references will be shown in the next issue.

計量方言学 (1) (要約)

井 上 史 雄

この論文では，方言分布データへの多変量解析法の適用の実例を略説し，また，重心法という単純な技法を紹介し，多くの語形の分布パターンを要約しうることを示す．

言語地理学的調査の分布データを計量的に扱う試みは，日本でも欧米でも様々なされてきた．計量的手法のうち，多くの語の等語線を1枚の図に描き，等語線の束を求めるという等語線重ね合わせ法は，古典的な手法といえる．似た分布パターンの語の使用度を地点毎に加算して図化する方法もある．またある地点の答を基準にして，一致度を計算する手法もある．算術的技法ともいうべき等語線重ね合わせ法や一致度計算は，電卓と時間と根気さえあれば手計算も可能である．それに対し，多変量解析法は，時間があったとしても手計算では到底不可能な解析法である．多変量解析法(または多次元解析法) は計量方言学の本領といってよい．

多変量解析法の実例として，数値データについて，いわゆる「河西データ」のクラスター分析・因子分析の適用結果をみる．また非数値データ (名目変数) については，共通語の文法形式への「林の数量化第3類」適用結果をみる．

多変量解析法の適用の結果，沢山の地点の沢山の単語を使って，一度に方言区画を試みることも可能になった．同時に，多量のデータを単純化し，パターン全体の内在的構造を取り出して観察できる．ひとたびパターン全体が把握できると，主要な特徴に注目することによって，簡単な算術的計算でも似た結果を得ることができる．