Dialectologia 1 (2008), 65-81.

GEOGRAPHICAL DISTANCE CENTER AND MULTIVARIATE ANALYSIS OF THE STANDARD JAPANESE[1]

Fumio Inoue

Meikai University, Japan

innowayf@yahoo.co.jp

Abstract

In this paper a new technique for representing dialectal differences will be introduced. A kind of simplification will be attempted to represent the distribution patterns of the lexical items of standard Japanese. In order to simplify the geographical distribution patterns, railway distance center graph is utilized. In this technique geographical locations are plotted on a one-dimensional line from a cultural center. The shift of the main cultural center of Japan from the west to the east is reflected in the graphs obtained from factor analysis and cluster analysis, and in the geographical distribution patterns of the standard Japanese words making use of the railway distance.

The process of dissemination has been concisely summarized by the railway distance center graph. Multivariate analysis allows us to grasp an overall picture of the relationship between dialectal distribution and the historical background of words. After applying multivariate techniques the results can be represented by more concise and simplified numerical techniques.

Key words

Distribution pattern, railway distance, factor analysis, cluster analysis, standard Japanese.

[1] This paper was presented at ICL 17 Prague 2003 (Inoue in print).

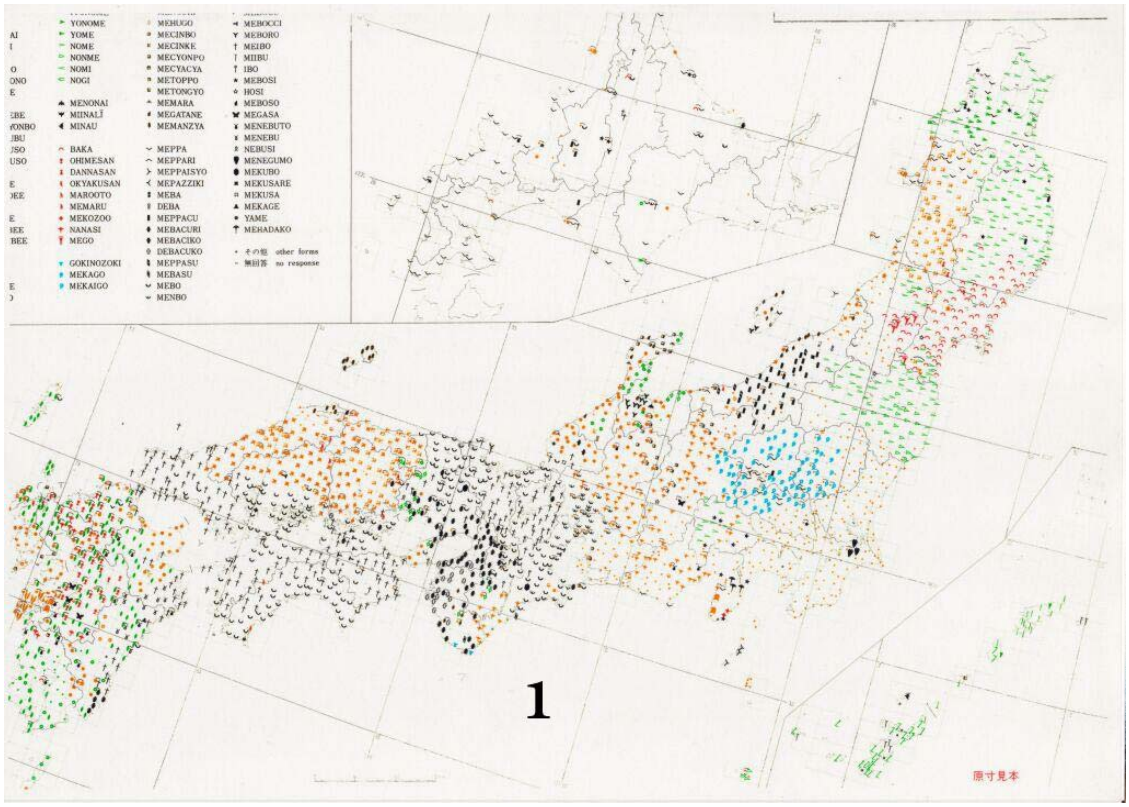
1. Theoretical and descriptive background

In this paper a new technique for representing dialectal differences will be introduced. A kind of simplification will be attempted to represent the geographical distribution patterns of the lexical items of standard Japanese (Inoue 2002). In this technique geographical locations are plotted on a one-dimensional line by making use of railway distances from a cultural center. The shift of the main cultural center of Japan from the west to the east (Inoue 2004) is reflected in the graphs obtained from factor analysis and cluster analysis, and in the geographical distribution patterns of the standard Japanese words.

2. Quantitative methodology

2.1. Kasai data: Basic Matrix of Words and Prefectures

The most reliable data on the dialectal distribution of Japanese is recorded in the "Linguistic Atlas of Japan" or LAJ (NLRI 1966-1974). Figure 1 is a sample map of LAJ with 2400 localities. There is also a database which shows the general tendencies of standardization of Japanese based on the lexical items of the "Linguistic Atlas of Japan". This computational data was compiled by Ms. Kasai and is referred to as the Kasai data.



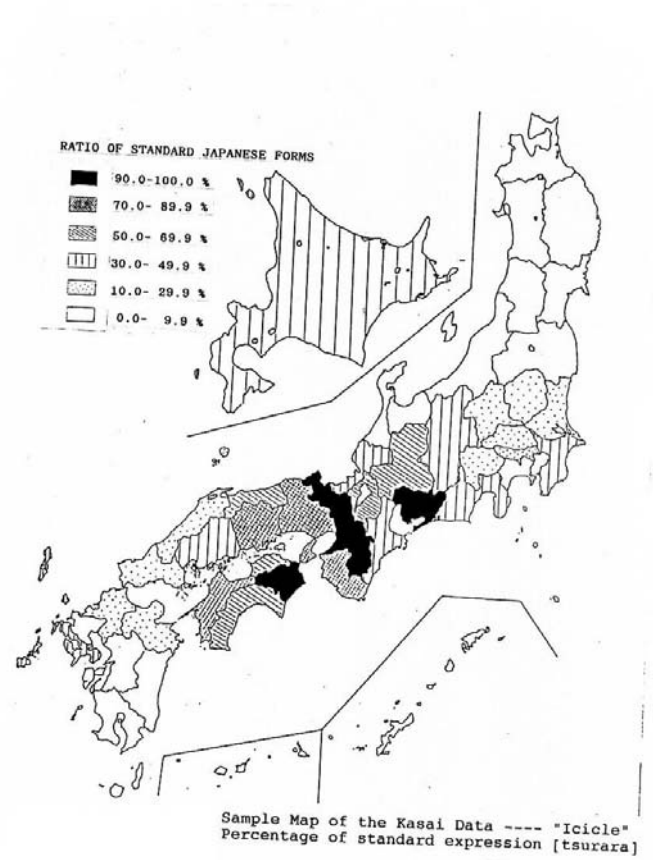
Map of the *Linguistic Atlas of Japan*

Figure 2 shows the basic structure of the data matrix of the Kasai data. For each of the 82 standard Japanese words the average degree of usage was calculated for each of the 48 prefecture areas.

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	...
:	:	:		:		:	:		:	
:	:	:		:		:	:		:	
:	:	:		:		:	:		:	

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

Figure 3 is a sample map of the Kasai data, showing percentage of standard Japanese words.



3

2.2. *Historical Background: the First Appearance*

More information was added to all the standard Japanese words of the Kasai data. For this presentation, information on usage derived from historical documents was utilized. The first appearance of the lexical items was determined on the basis of "Nihon Kokugo Daijiten" (*The Great Dictionary of the Japanese Language*) (2nd Edition) (Nihon Daijiten Kankokai 2000-2002), the largest Japanese dictionary at present, which cites historical examples and states the year that the word was first recorded in documents. There are problems in using data from a dictionary. One problem is that the first appearance of a word in a document may not necessarily coincide with the

first usage among people. However, as there is no other reliable source for this information, I have used the dictionary data in this study.

2.3. *Railway distances*

More information was added. The railway distances of the prefecture centers from both Kyoto and Tokyo were also added to the columns of prefectures.

Four factors to be considered

(1) As the geographical locations can be shown in one dimension by the railway distance center technique, the other dimension on a sheet of paper can be utilized for another factor.

(2) In this presentation the nationwide ratio of standardization is considered and a very interesting theoretical tendency was discovered.

(3) I have also analyzed the first appearance of the word-form in historical documentation.

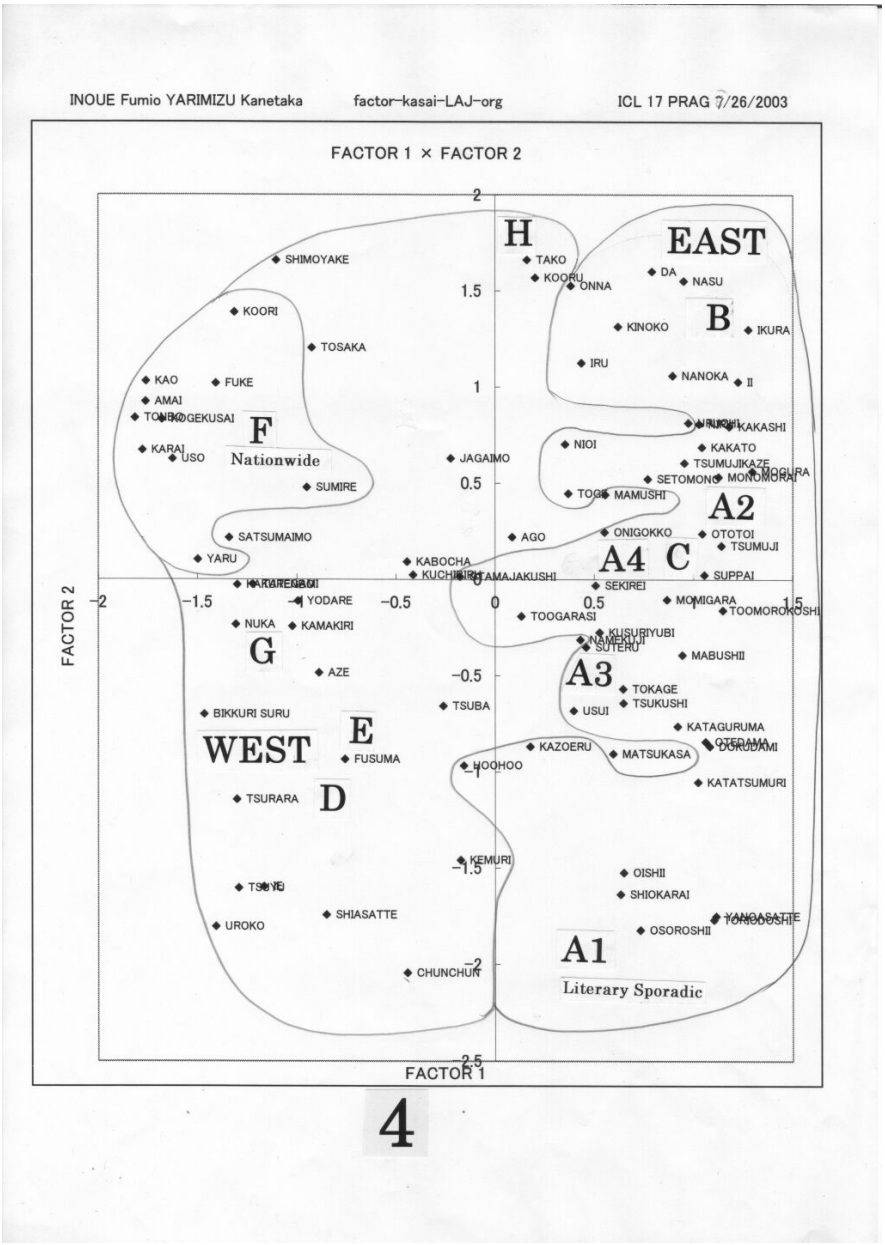
(4) The fourth phenomenon of the result of cluster analysis will be shown by symbols.

As there are four factors to be considered, a two-dimensional scattergram will be observed first. Following that, three-dimensional graphs will be shown in order to show the overall interrelationship of the three factors.

2.4. *Factor Analysis and Cluster Analysis*

In order to show the basic patterns of distribution of the Japanese standard words, the results of factor analysis of the Kasai data will first be introduced. The results of factor analysis and cluster analysis were most successful and coincided fairly well. Figure 4 shows the graph of factor load for

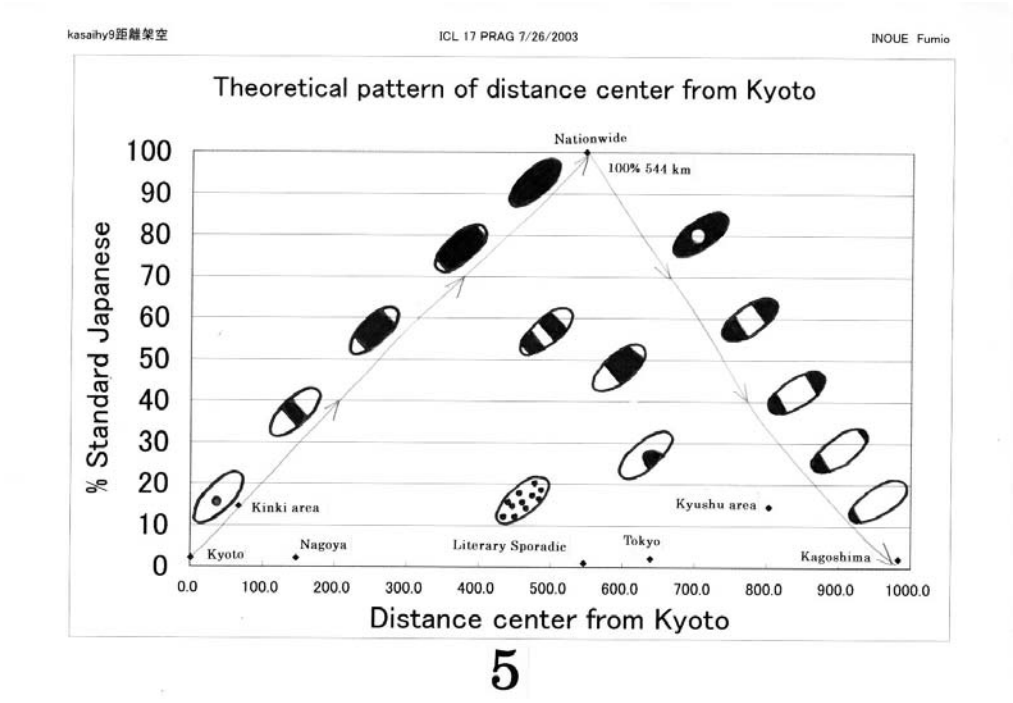
each of the 82 standard Japanese lexical items. In Fig 4, the results of the cluster analysis of the words are also entered by lines which encircle the words. The results of factor analysis and cluster analysis suggest that the present standard Japanese words are based both on western Kyoto factors and on eastern Tokyo factors. The shift of the main cultural center of Japan from west to east was found to be reflected in the graphs obtained from factor analysis and cluster analysis.



Results of the cluster analysis

3. Railway distance center

Now I will proceed to the newer portion of my research. A new technique of calculating the geographical locations of the 48 prefecture areas was adopted. It is the use of railway distance from cultural centers to the 48 prefecture centers. As the standard words were found to have disseminated from either Kyoto or Tokyo in the past, railway distance centers from Kyoto and Tokyo were calculated by making use of the railway distance from each of the prefecture centers to both Kyoto and Tokyo. Average railway distance centers were then calculated for each word in the same way as geographical gravity centers.



5

The formula of calculation for the railway distance center technique will be skipped in this paper. Instead, I will give concrete examples based on hypothetical language distribution data. In figure 5, the left to right axis or horizontal dimension shows the railway distance from Kyoto. The

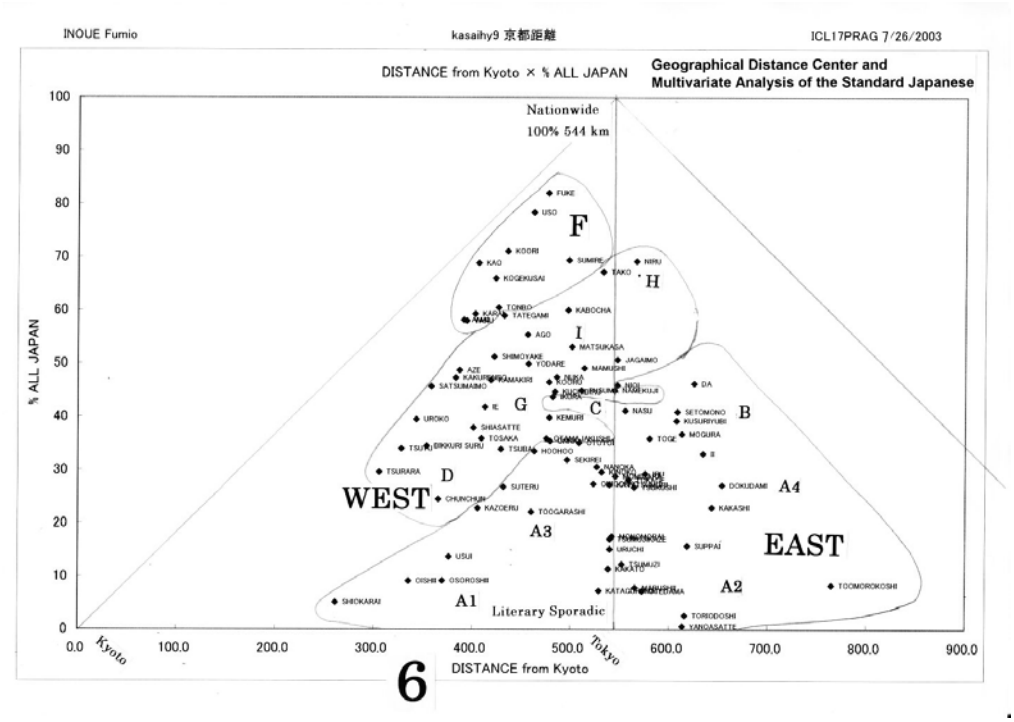
left-hand end of the graph corresponds to the location of Kyoto. The right-hand end corresponds to the farthest end of the Japan Islands. The bottom to top axis or vertical dimension shows the percentage of usage of the standard words for the whole of Japan. In Fig 5, hypothetical cases are calculated to show the pattern of diffusion of standard Japanese words. Small, simplified maps of Japan are added beside the curve. The dark areas of the maps represent the areas of distribution of the standard words.

The bottom left is the starting point of historical dissemination. If a word-form is used only in Kyoto prefecture, the railway distance center from Kyoto is 0 km, and usage ratio for the whole of Japan is about 3%. If the word-form diffuses further and is used widely in the Kinki area near Kyoto, the railway distance center from Kyoto becomes 50 km, and ratio for the whole of Japan becomes about 15%. If a word-form disseminates further and is used all over Japan, the railway distance center from Kyoto is 544 km, and the ratio for the whole of Japan is of course 100%. If a word-form becomes obsolete near the Kyoto area because of other newer dialectal expressions, and used only in the Kyushu area far southwest of Kyoto, the railway distance center from Kyoto becomes 800 km, and the ratio for the whole of Japan becomes about 15%. If the word-form is further restricted in distribution and used only in Kagoshima prefecture, the railway distance center from Kyoto is about 1000 km, and the ratio for the whole of Japan becomes about 2%.

Thus, as the line in Fig 5 shows, the route of typical diffusion of a word-form from Kyoto to the whole of Japan and next to periphery, progresses from left to right. This hypothetical pattern is a simplified linguistic diffusion model of the wave-theory or Wellentheorie by Johannes Schmitd. The actual data below can be compared with this pattern.

4. Results: Scattergram

I will limit my discussion to the results to Kyoto only because the railway distance from Kyoto showed a neater pattern.



4.1. *Figure 5 & figure 6 hypothetical diffusion pattern*

In figure 6, the horizontal axis shows the railway distance from Kyoto. The left-hand end of the graph corresponds to the location of Kyoto. Tokyo corresponds to a distance of about 510 km from Kyoto. The vertical axis shows the overall ratio of usage for the whole of Japan. This graph is the same as figure 5 shown earlier as a hypothetical graph. The pattern as a whole looks like a triangle. It shows that many standard words are distributed in various ways within hypothetically

possible routes of diffusion from Kyoto. But the left-hand side of this graph is vacant showing that no standard words were found which are distributed near Kyoto. This fact tells that Kyoto has ceased to be the center of standard words at present. Also the right-hand rim of this graph is vacant showing that no standard words were found which are distributed only in the further end of Japan. The peak of this triangle is about 500 km away from Kyoto, and approximately coincides with the location of Tokyo, not with the location of nation-wide words.

4.2. *Figure 6 & figure 4 cluster analysis*

The result of cluster analysis is added by solid lines encircling clusters. The overall distribution of words in figure 6 is very similar to the result of cluster analysis in figure 4 especially when held diagonally. I have drawn various graphs based on arithmetic calculations since I had made the graph of cluster analysis. Fig 6 shows the best fit. The coincidence of distance center with the results of cluster analysis looks better than for factor analysis. There are fewer exceptional words in this distance center graph. The geographical distribution seems to have been successfully summarized with this rather simple arithmetic calculation.

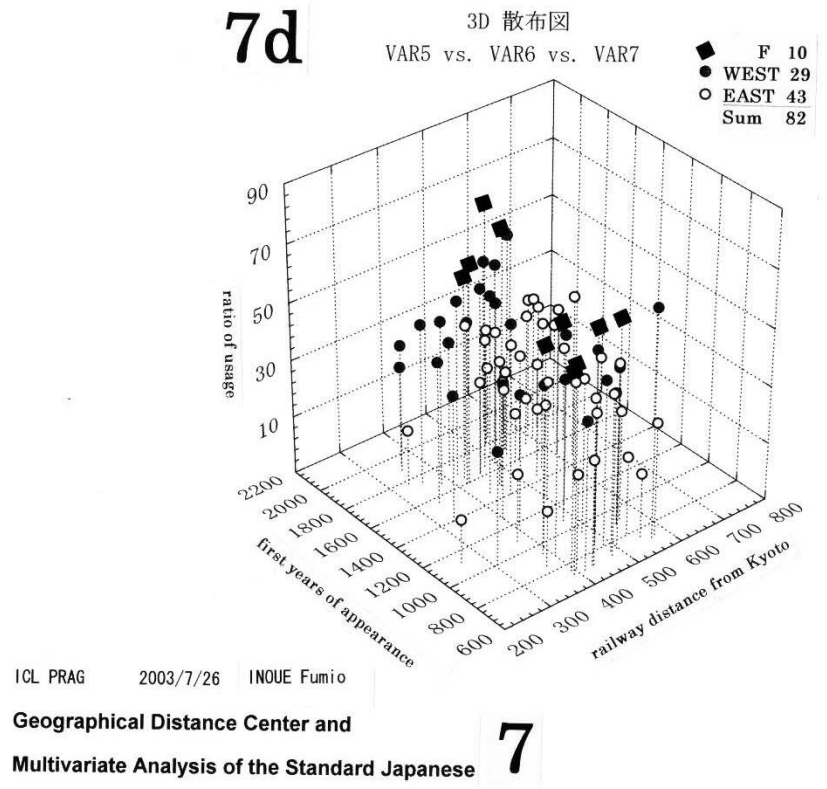
4.3. *Figure 6: 3 strata*

Three diagonal strata are observable in figure 6: the east cluster at the bottom, the west cluster above it, and the F cluster on the top. This positioning can be explained as follows. The words of the F cluster, which is originally a part of the west cluster, can be treated separately because of their compact distribution on the top of the triangle. Their centers are limited in geographical positioning, and they are used vigorously all over Japan, with all the words having a usage of over 50%. The

words of the west cluster disperse between 300 km and 600 km from Kyoto, and are used considerably all over Japan. The approximate center of usage is about 50%. The centers of the east cluster words are diffused rather widely (between 300 km and 800 km from Kyoto,) but are not used so much for Japan as a whole. There is a mass of distance centers near Tokyo with a usage of about 30%. These words are used near Tokyo and more widely in eastern Japan. This orderly positioning will later be explained from a historical perspective.

4.4. *Figure 6 & history*

The pattern of distribution of Japanese standard words seems to have been ruled by a geographic tendency which is mainly controlled by the distance from Kyoto and by the overall ratio of diffusion for all Japan. As the peak of the east cluster is near Tokyo, and as the analysis of the railway distance center from Tokyo has shown, some of the words in the east cluster were adopted as standard later in Japanese language history. These tendencies indicate that lexical items diffuse in a certain velocity from a cultural center, if taken as a whole. Thus, some historical background is necessary in order to interpret the patterns of the above data more accurately. The two-dimensional data of figure 6 above corresponds to the original matrix of the Kasai data. In computerizing the data, other kinds of information were added (Inoue 2004). One type of information is the first year of appearance of the word-form in historical documents. The first year of appearance was looked at next, but clear correlation with the results of cluster analysis was not found. This is partly natural because no historical information was included in the multivariate analysis above. However, it also suggests that Kyoto has been the hidden origin of standard Japanese for a long time (Inoue in press).

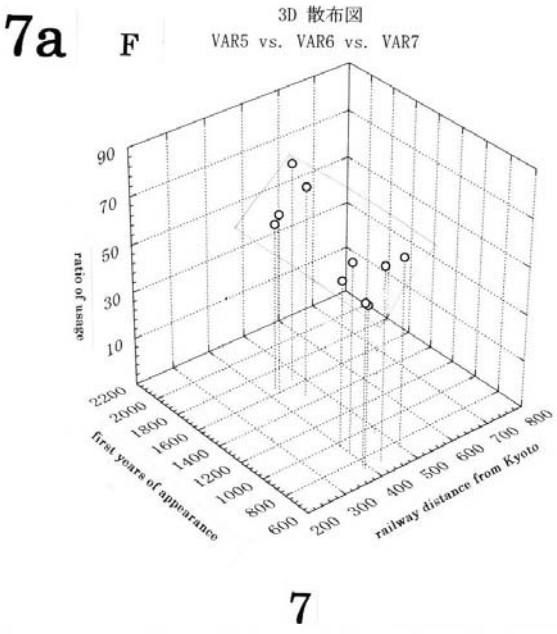


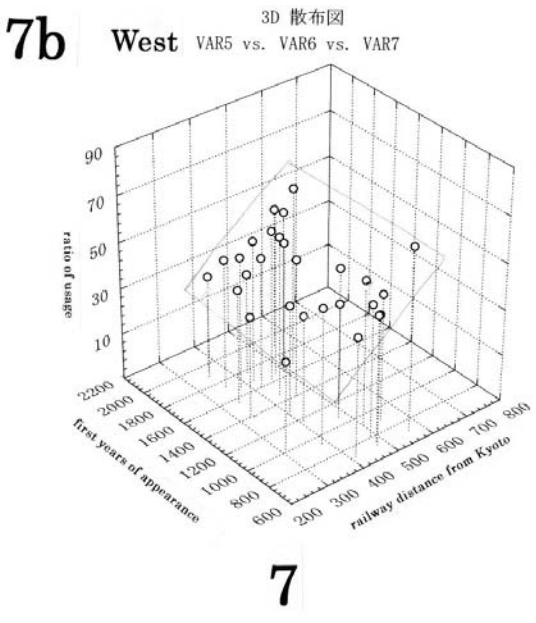
5. Results: 3D graph

In order to grasp overall patterns, three-dimensional graphs making use of ratio of usage, railway distance center and year of appearance, was drawn. The upper-left face of the three-dimensional graph in figure 7d corresponds to figure 6, the horizontal axis showing railway distance from Kyoto, and the vertical axis showing the overall rate of usage for the whole of Japan. The dimension on the base is a newer one. This shows the first year of appearance in a historical document. The years actually extend from the 8th century when the Japanese language was widely

recorded, to the 20th century. Three distinctions of clusters are shown by symbols. Thus, four kinds of information can be observed together in this graph. However, as there are as many as 82 balls in the graph, it is rather difficult to interpret the distribution pattern. So, three three-dimensional graphs will be shown separately for the three clusters.

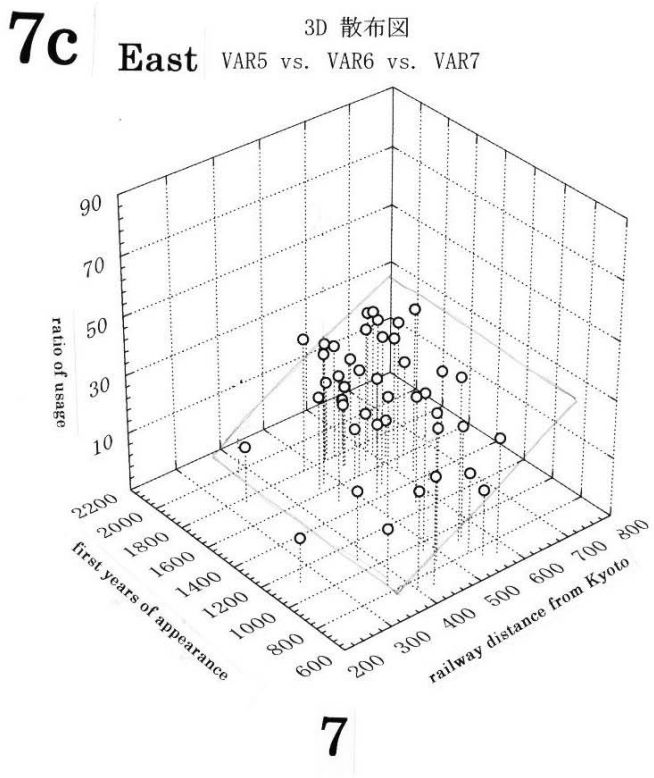
F cluster in Figure 7a is situated on the top of the other words with long legs, being used much all over Japan.





Western cluster words in figure 7b are situated in the center with legs of middle length, and show a cline from left to right.

The eastern cluster in figure 7c is situated at the lower layer of the words. The words of the eastern cluster also show a tendency to be distributed from the lower-left to the upper-right, indicating that recent words first cited in the 19th and 20th centuries are distributed mainly near Tokyo. These three graphs showed the relation between the three-layered pattern of geographical distribution and historical documents. The most prominent tendency is that most of the words in the eastern cluster appeared late in history.



After looking at the graphs separately, it is easier to see the general pattern of distribution from the graph of all the 82 words in figure 7d. In Fig 7d, F cluster words are shown by black

squares. In this graph, the two groups of F cluster are conspicuous at the top. The words of the western cluster, shown by black balls are concentrated on the left side, showing that they originated after the Middle Ages and have not attained the status of nationwide distribution. The words in the eastern cluster are shown by white balls. The words in the eastern cluster that appeared early in history, are distributed separately from the other words. The eastern words at the low back corner of this graph, with a recent history and a distant distribution from Kyoto, can also be grasped as a group, although half of them are covered by western cluster words. More recent words appearing after the 15th century expanded from central Japan to the east in the direction of Tokyo. This seems to show that recent standard words were first adopted in Tokyo and disseminated near Tokyo.

6. Conclusions

6.1 *Railway Distance Center and History*

As shown in the hypothetical curve in figure 5, the standard words must have diffused from Kyoto in the past, disseminating further and further as time passed. However, in Japanese history a new cultural center appeared later and became the new capital of Tokyo in eastern Japan (Kawaguchi and Inoue 2002). Because of the new wave of standardization from Tokyo, the eastern cluster appeared and diffused from Tokyo, so that the word-forms belonging to the western cluster also added power because of this standardization. In Figure 7d, F cluster must have been accumulated on the top of the western cluster by acquiring power also from Tokyo.

In conclusion, F cluster is the cluster of words gathering western and eastern tendencies of standardization. The western cluster includes words which disseminated from Kyoto. The eastern cluster includes words which are characteristically used in eastern prefectures. Thus, three strata can

be interpreted historically. Although the graph of Fig 7d seems to show clear distribution patterns, it is not due to the power of Kyoto itself, but due to the recent powerful dissemination from Tokyo. Recent waves from Tokyo were again observed in the railway distance center graph. In summary, the railway distance center method showed quite successfully the relationship with the first year of appearance of the standard words (Inoue 2006).

6.2. *The theoretical implications of the Railway distance center Method*

In order to simplify the geographical distribution patterns, railway distance center graphs were drawn. The process of dissemination has been more concisely summarized by the railway distance center graph (Inoue in press). Multivariate analysis allows us to grasp an overall picture of the relationship between dialectal distribution and the historical background of words. After applying multivariate techniques the results can be represented by more concise and simplified numerical techniques (Inoue 1996a, b). Representation in two dimensions using railway distance center method can be applied to any geographical distribution in any country if the center of linguistic diffusion is clear. As this simple technique has proved interesting for Japanese, I would like to apply the same technique to the dialects of other languages. Although the dialectal data is from within one country, the methodology and technique of analysis is international. This International Journal on the internet is the most appropriate place to exchange ideas of analysis.

References

INOUE, Fumio (1996a, 1996b) "Computational dialectology (1) (2)", *Area and Culture Studies*, 52, 67-102, 53, 115-134.

INOUE, Fumio (2002) "Dialect gravity center and rate of usage of standard Japanese forms", *Area and Culture Studies*, 63, 115-121.

INOUE, Fumio (2004) "Multivariate analysis, geographical gravity centers and the history of the standard Japanese forms", *Area and Culture Studies* 68, 15-36.

INOUE, Fumio (2006) "Geographical Distance Center and Rate of Diffusion of Standard Japanese", in *Proceedings of the 4th International Congress of Dialectologists and Geolinguists*, ed. A. Timuska. Riga, Latvia, 239-247.

INOUE, Fumio (in print) "Geographical Distance Center and Multivariate Analysis of the Standard Japanese", in *Proceedings of the International Congress of Linguists 17* (Prague) (CD-ROM).

INOUE, Fumio (in press) "Abstracting Geographical Space by Railway Distance: Year of attestation and diffusion of Japanese dialects", Lameli *et al.* (ed.), *Sprachraum and infrastructure* (Mouton de Gruyter).

KAWAGUCHI, Yuji and Fumio INOUE (2002) "Dialectology in Japan in historical perspectives, Pt. II: Historical characteristics and geographical distribution of standard Japanese forms", *Revue Belge de Philologie et d'Histoire*, 80, 816-829.

Nihon Daijiten Kankokai (2000-2002) *Nihon Kokugo Daijiten* (Great Dictionary of the Japanese Language), Second edition (Shogakukan).

NLRI (National Language Research Institute) (1966-1974) *Linguistic Atlas of Japan* Vols. 1-6 (Okurasho Insatsukyoku).