Finding representative sets of dialect words for geographical regions

Marko Salmenkivi

Helsinki Institute for Information Technology, Basic Research Unit
Department of Computer Science
P.O.Box 68
FIN-00014 University of Helsinki
Finland
marko.salmenkivi@cs.helsinki.fi

Abstract

We investigate a corpus of geographical distributions of 17,126 Finnish dialect words. Our goal is to automatically find sets of words characteristic to geographical regions. Though our approach is related to the problem of dividing the investigation area into linguistically (and geographically) relatively coherent dialect regions, we do not aim at constructing more or less questionable dialect regions. Instead, we let the boundaries of the regions overlap to get insight to the degree of lexical change between adjacent areas. More concretely, we study the applicability of data clustering approaches to find sets of words with tight spatial distributions, and to cluster the extracted distributions according to their distribution areas. The extracted words belonging to the same cluster can then be utilized as a means to characterize the lexicon of the region. We also automatically pick up words with occurrences appearing in two or more areas that are geographically far from each other. These words may give valuable insight to, e.g., the study of cultural history and history of settlement.

1. Introduction

In this paper we investigate a corpus of geographical distributions of 17,126 Finnish dialect words. The data are based on a small part of the Lexical Archives of the Finnish Dialects in the Research Institute for the Languages of Finland. The project for a comprehensive dictionary of Finnish dialects is in progress, and in connection with the dictionary project the regional distributions of a large set of dialect words have been stored in electronic form. These data comprise the corpus studied in this paper. Along with the proceeding of the dictionary project, the number of digital maps increases. Thus, data analysis and data mining methods are needed for extracting knowledge automatically from the corpus.

Our goal is to automatically find sets of words characteristic to geographical regions. Identifying dialect regions is more or less a question of interpretation. Inside such a region the linguistic variation should be relatively slight compared with the variation across distinct areas. The language variation is gradual, and even if border lines can be found, different linguistic features do not necessarily follow the same lines (see, e.g., Heeringa and Nerbonne, 2002; Nerbonne and Kretzschmar, 2003).

Though our approach is related to the problem of dividing the investigation area into linguistically (and geographically) relatively coherent dialect regions, the goal is not to construct dialect regions. Instead, we let the boundaries of regions overlap, and want to find sets of words characteristic to different regions. In this way we wish to get insight to the degree of lexical variation in adjacent areas.

In this paper we present results of applying relatively simple data clustering approaches to the following problems:

1. Finding sets of words with tight spatial distributions. By tightness we mean that a word is used in a set of municipalities that are located geographically close to each other, and it is not used elsewhere.

2. Clustering the extracted distributions according to their distribution areas. The extracted words belonging to the same cluster can then be utilized as a means to characterize the lexicon of the region.

3. Finding words appearing in two or more areas that are geographically far from each other, and isolated in the sense that there are no occurrences between the areas. Explanations for such distributions of occurrences need to be given by linguistic experts; the dialect words may give valuable insight to the study of cultural history and history of settlement, for instance.

This paper is organized as follows. We first introduce the related work in Section 2, and the corpus in Section 3. In Section 4 the methods for extracting and clustering interesting distributions are described. Section 5 summarizes the results, and Section 6 takes a look at ongoing and future work. Section 7 is a brief conclusion.

2. Related work

Dialectometry means measuring linguistic differences primarily with respect to geography. While many of the early studies in the field tried to identify dialect regions, the later research has mainly concentrated on the different aspects and the continua of the complex phenomenon of linguistic variation (Nerbonne and Kretzschmar, 2003). Nerbonne and Kleiweg (2003) introduce a lexical distance measure, and analyze quantitatively the lexical variation of the sites of the Linguistic Atlas of the Middle and South Atlantic States. Palander et al. (2003) apply hierarchical clustering in their detailed analysis of Finnish linguistic variation based on morphological and phonological features of 198 idiolects from Savonlinna region in Southeast Finland. Leino et al. (2006), and Hyvönen et al. (2006) investigate several multivariate methods, e.g., principal components analysis and clustering, the aim being to summarize the distributions of 9,600 Finnish dialect words. Their corpus includes approx. 56 % of the word distributions analyzed in this paper. The traditional view of Finnish dialect

regions is mainly based on Kettunen (1940). The Finnish dialect regions as described by Savijärvi and Yli-Luukko (1994) are shown in Fig. 4 below. Embleton and Wheeler (1997, 2000) create machine-readable forms of Kettunen's dialect atlas, and apply multidimensional scaling to the data analysis. For further references on the study of Finnish dialects, see, e.g., Palander et al. (2003). See also Nerbonne and Kretzschmar (2003) for further references to computational techniques in dialectometry.

3. Corpus

The project for a comprehensive dictionary of Finnish dialects is in progress in the Research Institute for the Languages of Finland. This huge work will eventually consist of 20 volumes, and it is expected to be finished in the 2030's. Seven volumes have been completed, and five of them are available in SGML-format. The total number of data items in the Lexical Archives of the Finnish Dialects is approx. 7 million word instances, mostly stored in a card index (Tuomi, 1989). The beginning of collecting these data goes back to the 19th century. The project was established in 1896, and the latest data are from the 1970s. In the very beginning the aim was to gather the Finnish lexicon from the whole country exhaustively. As this goal turned to be impossible to reach the methods changed and developed. In the 1920s the country was divided into 23 regions; the intention was to describe the lexicon of at least one municipality from each region completely. In addition to educated recorders in the selected municipalities, there were a lot of voluntary correspondents – at least one correspondent in almost every municipality – that significantly contributed to the project (Tuomi, 1989). Nevertheless, the research activity varies remarkably between municipalities. The dictionary project has produced a large number of maps describing the distribution areas of dialect words, and most of them have now been stored in electronic form. The corpus studied in this paper was constructed based on these maps. In the corpus a set of municipalities is associated with each dialect word; the word is known to be used in those municipalities. (More precisely, in some cases there are different phonetic variants of the same "word". However, we use the inaccurate term "dialect word" below). As ancillary information we use the knowledge of the geographical neighbours of each municipality as well as the distances between the municipalities.

The total number of dialect word instances, that is, word-municipality pairs in our corpus is 391,180. This is a small fraction of the data in the Lexical Archives of the Finnish Dialects. A major reason is that during the dictionary project words are mainly processed in alphabetical order, and the project is unfinished. Our corpus includes mostly words of which initial letter is between A and K. This fact causes some regional unbalance; for instance, letter F strongly refers to the influence of Swedish, and thus, the westernmost parts of Finland. It should be noted that a very large part of the dialect words in the archives consists of very local, or rare words, occurring typically in only one or two municipalities. These words are not included in our corpus either. Similarly, no maps of the most common words are available, for the practical reason that the distributions cover the whole country. Thus, the words in our corpus are between these extremes, and, it is plausible to expect that they are the most informative ones when it comes to the characterization of regional lexicons.

The total number of municipalities inside the present borders of Finland is 448. In addition, the corpus includes occurrences from 75 other municipalities and a few larger areas, mainly from the municipalities that Finland ceded to the Soviet Union after the Second World War. In this paper we investigate only the municipalities of Finland in the present time. Later the analysis should be extended to the other areas as well.

There are 27 municipalities with more than 2,000 occurrences, and 132 municipalities with more than 1,000 words (see Figure 1, top). Furthermore, there are 67 municipalities with no recorded dialect words in the set of 17,126 words. Most of the zero-municipalities are, however, in the Swedish-speaking districts of Finland, and, thus, can be ignored when analysing Finnish dialects. The bottom panel of Figure 1 indicates the proportion of words (x-axes) that occur in at most the number of municipalities indicated by the y-axes. We see, for instance, that more than 50 % of the word distributions cover less than 10 municipalities, and 90 % of them cover at most 50 municipalities. Thus, a majority of the words are local in the sense that they only occur in a small fraction of all the municipalities.

4. Method

We are interested in finding sets of words characteristic to geographical regions. To reach the goal we should, on one hand, assess the similarity between two specific occurrences (municipalities) of a dialect word, and, on the other hand, the similarity between the distributions (that is, sets of occurrences) of two dialect words.

In Finland the areas of municipalities vary remarkably in different parts of the country, reflecting mainly the differences in the density of population. In the north, the municipalities are very large compared to those in the southern part of the country. Thus, the absolute distances between municipalities may be very different in different regions.

As the measure of the pairwise distance between two occurrences of a single dialect word, we employed the length of the path in the adjacency graph of the municipalities. Then we applied the following two-phase clustering algorithm. In the first phase, we clustered the occurrences of each word by using a simple depth-first search algorithm. Starting from an arbitrary occurrence the algorithm searches for another occurrence in an adjacent municipality. In the case such an occurrence is found, it is assigned to the same cluster as the previous one, and the search is conducted recursively. If no occurrences are found, the current cluster is completed. A new starting point is selected from the set of the occurrences not yet assigned to a cluster. The procedure continues until every occurrence is assigned to a cluster. As the result, the occurrences of the word are divided into one or several clusters, roughly corresponding to the different regions where the word is used.

Evidently, the approach can be generalized to allow one or more skips in a path in the neighbourhood graph, that is, municipalities without an occurrence. The skips can be

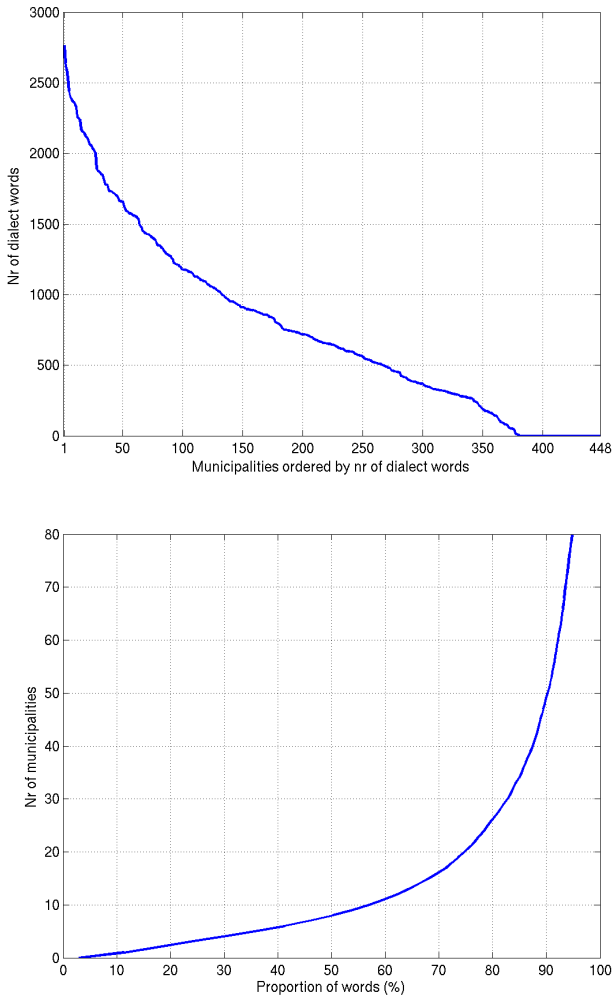


Figure 1: Municipalities ordered in decreasing order by the number of words in our corpus (top). Proportion of words (x-axes) occurring in at most the number of municipalities indicated by the y-axes (bottom).

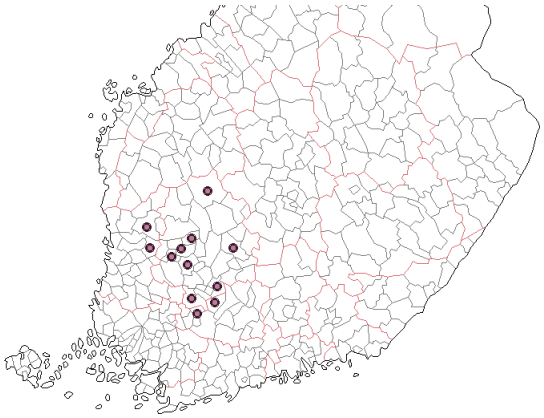


Figure 2: Example of a tight distribution (*ensipoikainen*), one skip allowed in the adjacency graph.

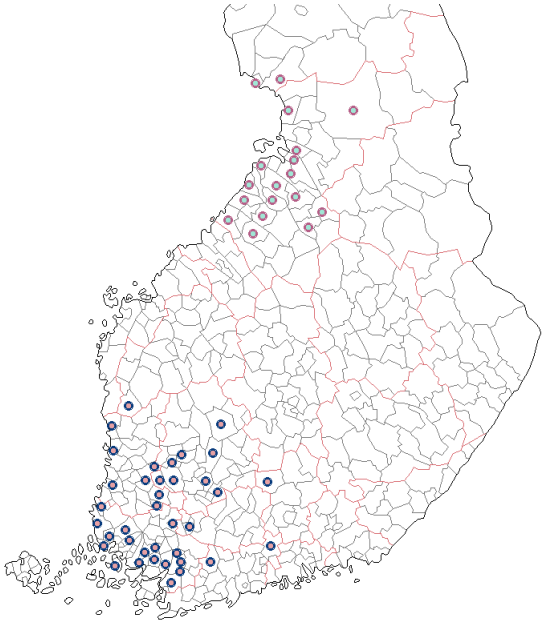


Figure 3: Example of a distribution with two geographically isolated distribution areas (*aanata* 'anticipate, foresee').

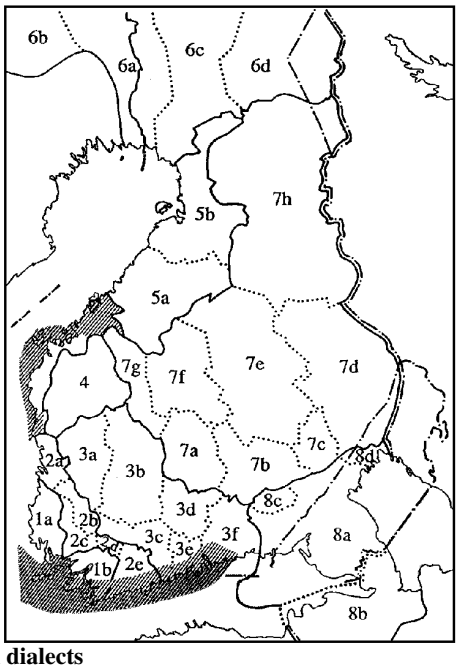
used as a simple way of handling the problem of incomplete data (see examples in Figures 2 and 3).

To find clusters of words with tight distributions in the same regions, those words whose occurrences consist of only one cluster are selected for the second phase. In the second phase the hierarchical clustering algorithm is employed, and the words are clustered with respect to the similarity of the distributions of their occurrences. We defined the distance of two distributions of dialect words as the average of the pairwise distances between the occurrences of the words, that is, $dist(w_1, w_2) = \sum_{i=1}^{k} \sum_{j=1}^{n} d(m_i, m_j^{'})/kn$. where w_1 and w_2 are words, the municipalities of occurrences being $m_1, \ldots, m_k$, and $m_1^{'}, \ldots, m_n^{'}$, respectively, and $d(m_i, m_j^{'})$ is the distance between the occurrences m_i of w_1, and $m_j^{'}$ of w_2. Here we used the Euclidean distances between the occurrences.

The first-phase algorithm can also be employed to find words with geographically isolated regions of occurrences. In that case two occurrences are assigned to the same cluster unless there are at least $k >> 0$ successive municipalities without occurrences between them. Then, if several clusters are found, they are isolated from each other. If needed, the resulted occurrences in each region can be further clustered to evaluate, for instance, whether they are sufficiently tight for the purposes of the current analysis.

5. Results

We ran the described clustering algorithm on the dialect word corpus to extract the tight distributions. As the criteria for choosing a word a into the set of tight distributions we employed the following requirements: first, the set of occurrences M_a must include at least 4 municipalities, and no more than 20 municipalities. This was to guarantee that the distributions are sufficiently local. Second, in the adjacency graph of all the 448 municipalities there has to be such a path between any pair $(i, j), i, j \in M_a$, that for all



Western dialects

1. South-Western dialects
2. Mid-South-Western dialects
3. Tavastian dialects
4. Southern Ostrobothnian dialects
5. Central and Northern Ostrobothnian dialects
6. Northernmost dialects

Eastern dialects

7. Savonian dialects
8. South-Eastern dialects

Figure 4: Finnish dialects (Savijärvi and Yli-Luukko, 1994)

the successive municipalities $k, l \in M$ in the path it holds that either $k \in M_a$, or $l \in M_a$ (or both). Here M is the set of all the municipalities. In other words, in the path single municipalities with no occurrence are allowed but two successive municipalities with no occurrence are not allowed. Applying these criteria on the corpus resulted in 1012 distributions (6 % of total). An example of a distribution satisfying the requirements is given in Figure 2. Of course we also conducted trials with other parameter values with slightly different results. Allowing several skips in a path in the neighbourhood graph rapidly extends the set of distributions.

In the second phase we clustered the selected distributions by the hierarchical clustering algorithm, and stopped the clustering when 100 clusters were left. Most of the clusters consisted of single words. Four of the resulting clusters included more than 100 words, and 11 of them consisted of more than 20 words. Figures 5 and 6 illustrate the covers of the geographical regions of the largest clusters. The greater the circle drawn on the municipality, the greater the proportion of the words of the cluster that occur in the municipality.

For comparison of the results, we also present a division of Finnish dialects into eight regions in Fig. 4, the division being mainly based on phonological and morphological features (Savijärvi and Yli-Luukko, 1994).

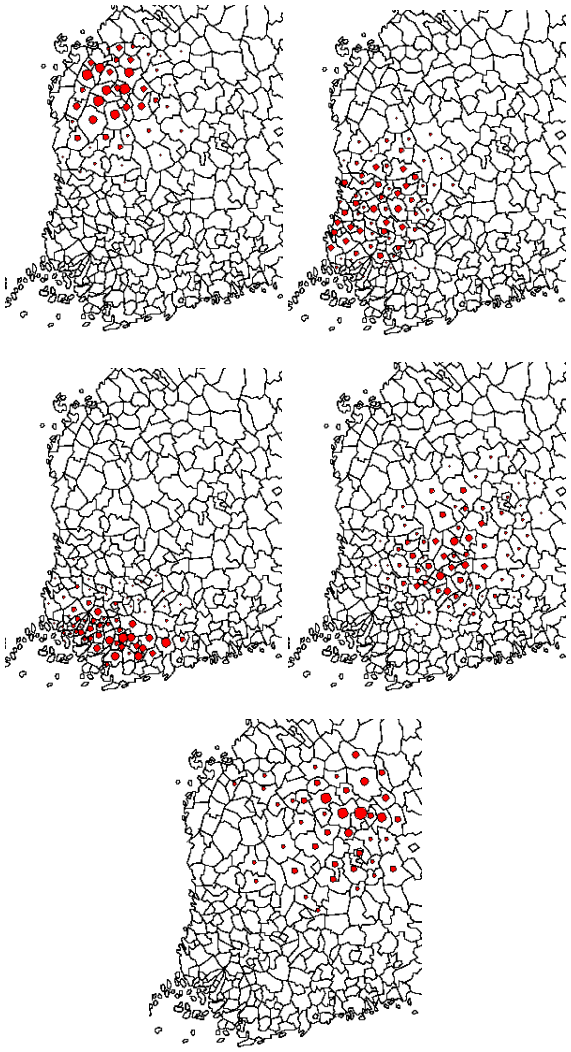


Figure 5: Summary of clustering the tight distributions, West Finland.

The largest cluster in the sense of number of dialect words (237 words, Figure 5, top-left) agrees very closely on the area of Southern Osthrobothnian dialects in Fig. 4. There is a relatively large set of core municipalities that cover remarkable proportions of the dialect words in the cluster. Hence, a large set of representative words can be assigned to the district. The distinctiveness of the Southern Osthrobothnian dialects has been noticed in earlier studies as well. Still, the gradual change of lexicon is indicated by the fact that there are municipalities from dialect regions 3a and 7g that also share part of the word distributions in the cluster.

The clusters depicted on the top-right (133 words), centre-left (45 words), and centre-right (101 words) panels are located across the regions of South-Western, Mid-South-Western and Tavastian dialects. In the top-right and centre-right clusters, the variation between the word distributions is relatively large, indicated by the small sizes of the circles. In the small cluster depicted on the bottom panel, a few core municipalities in the Savonian region dominate

the distributions but the individual word distributions scatter to the Southern Osthrobothnian and Tavastian regions.

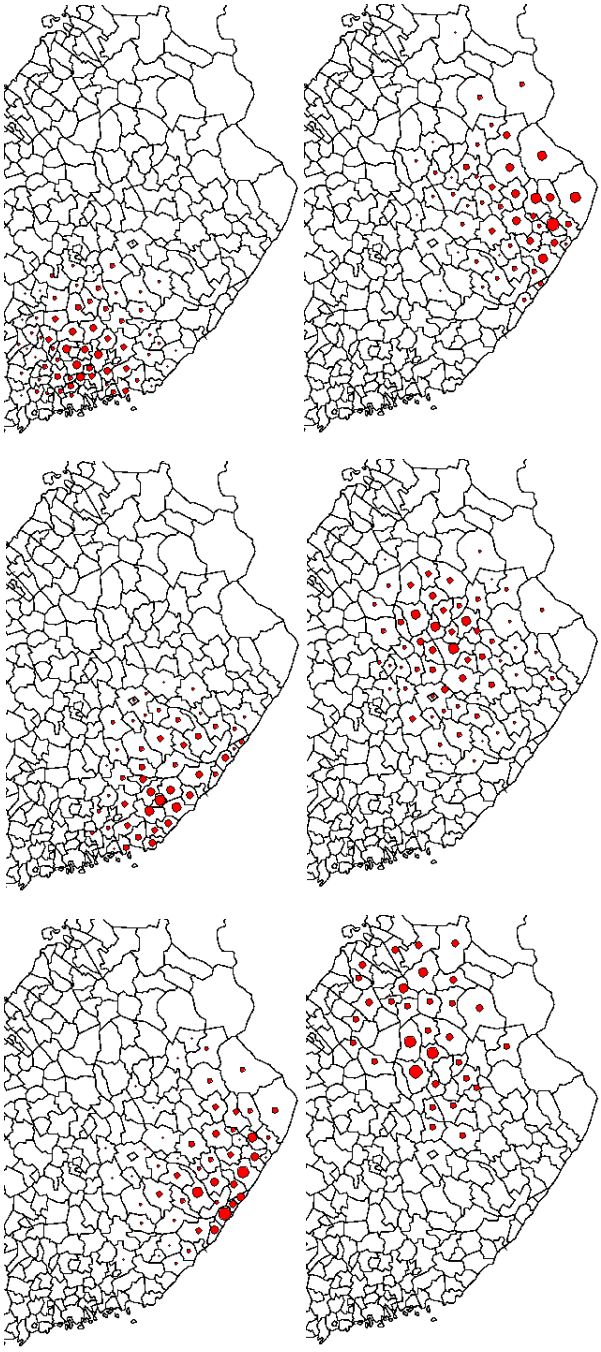


Figure 6: Summary of clustering the tight distributions, East Finland.

The Savonian dialect region is large, and all the clusters summarized in Fig. 6 include word distributions that cover municipalities in the region. The gradual change of lexicon between the regions of Fig. 4 is demonstrated by several of the clusters in Fig. 6. The largest of the clusters (104 words, top-left) settles down in the south-east, the emphasis being on the Tavastian and Savonian dialect regions, only very slightly reaching the southernmost municipalities in the province of Karelia, and the region of the South-Eastern dialects. A kind of counterpart is the centre-left cluster (58 words) that is located around the town of Lappeenranta (area 8c in Fig. 4), and thus, has its geographical emphasis in South Karelia. The bottom-left cluster of 28 words covers municipalities further to the north-east, in the South-Eastern as well as the Savonian regions. The right-side clusters (top: 59 words, centre: 57 words) are located in the Savonian region; the small cluster of 5 words on the bottom-right also reaches the Northern Osthrobothnian region.

The clusters in Fig. 7 include the distributions of 27 (top), 32 (centre), and 15 (bottom) words. The regions are mainly Savonian (top), Central and Northern Osthrobothnian (centre), and Central/Northern Osthrobothnian and Savonian (bottom).

The rest of the words were assigned to clusters of only 1–5 word distributions. The total of 908 words (90 % of all the selected distributions) belonged to the clusters summarized in Figures 5–7.

The small sizes of the clusters in the north and north-east apparently reflect the use of pairwise Euclidean distances when employing the hierarchical clustering algorithm in the second clustering phase. Hence, further experiments with the shortest-path distance measure could reveal more when it comes to the distributions in the north.

In the case of isolated regions of occurrences, the first-phase algorithm could be used to easily find distributions with several geographically distinct areas. For instance, when requiring two separate areas that could not be reached with less than seven steps in the neighbourhood matrix, and with at least three occurrences in both areas, 857 distributions were selected. (An example was given in Fig. 3). Clustering these distributions is apparently much harder than in the case of tight distributions. We clustered each of the distinct subdistributions of a dialect word separately, and, finally, we assigned two words w_1 and w_2 to the same cluster, if for each separate set of occurrences of w_1, one of the separate sets of occurrences of w_2 belonged to the same cluster after the second-phase clustering. This practice resulted in some interesting clusters but in general more sophisticated methods should be devised.

6. Ongoing and future work

The variation of research grade in different municipalities has several reasons. One of the most important is the decision made in the 1920s: the country was divided into 23 districts, and a sample set of municipalities – at least one from each district – was selected. The dialect words were intensively recorded in those municipalities. While this approach made it possible to collect almost the whole lexicon from some municipalities, it has some problems when the data are investigated quantitatively. Relatively large amount of data from a set of municipalities that were selected based on the knowledge of dialects (and not only lexical knowledge) of that time may bias the analysis to some extent. For these reasons, if the lexical variation was to be studied quantitatively based on solely lexical data, the sampling process of the municipalities should be taken into account. A closely related problem is the incompleteness of the data in many municipalities.

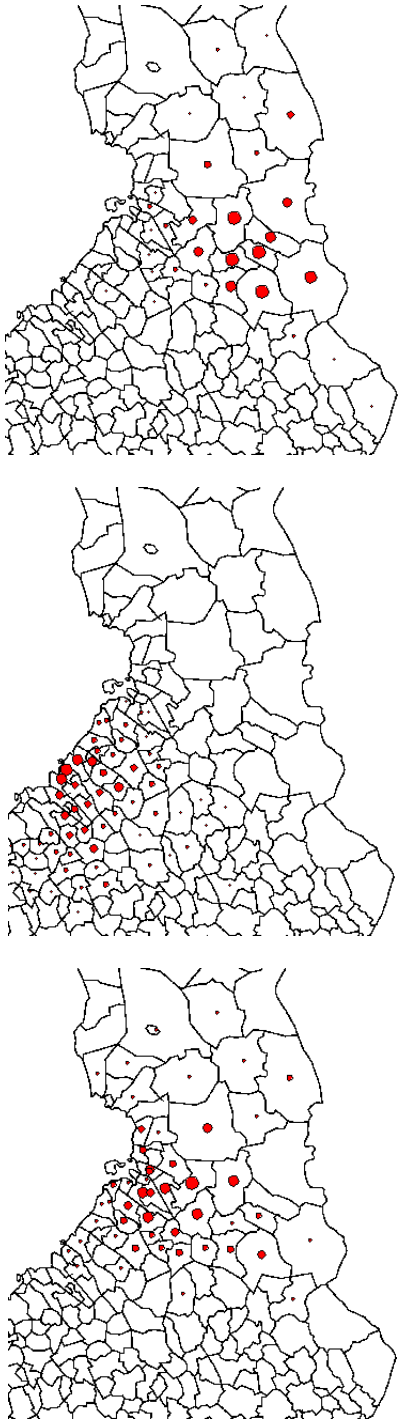


Figure 7: Summary of clustering the tight distributions, Central/North Finland.

We are currently modeling the sampling process of the municipalities and the missing data by Bayesian hierarchical modeling with spatial (Markov random field) dependencies, and implementing efficient Markov chain Monte Carlo simulators for estimating the model parameters. The goal is to evaluate whether modifying the data based on the model influences the observed linguistic variation. One of the interesting features of such models is the possibility of taking ancillary information, such as known water routes, historical borders etc. into account when modeling the spatial interaction between areas.

7. Conclusion

In connection with the project for a comprehensive dictionary of Finnish dialects geographical distributions of a large set of dialect words have been stored in electronic form. These 17,126 distributions comprise the corpus studied in this paper. We demonstrate how simple clustering algorithms can be used to extract representative sets of dialect words for different regions. More complex clustering methods are needed for automatically clustering words appearing in two or more areas that are geographically far from each other, and isolated in the sense that there are no occurrences between the areas.

8. References

S. Embleton and E. S. Wheeler. 1997. Finnish dialect atlas for quantitative studies. *Journal of Quantitative Linguistic* 4 (1–3):99–102.

W. Heeringa and J. Nerbonne. 2002. Dialect areas and dialect continua. *Language Variation and Change* 13:375—398.

S. Hyvönen, A. Leino, and M. Salmenkivi. 2006. Multivariate analysis of dialect data. Manuscript submitted to *Literary and Linguistic Computing* (under review process).

L. Kettunen. 1940. *Suomen murteet III A. Murrekartasto*. Suomalaisen Kirjallisuuden Seura.

J. Nerbonne, and W. Kretzschmar 2003. Introducing Computational Techniques in Dialectometry. *Computers and the Humanities*, 37:245–255.

J. Nerbonne, and P. Kleiweg. 2003. Lexical Distance in LAMSAS. *Computers and the Humanities*, 37:339–357.

A. Leino, S. Hyvönen, and M. Salmenkivi. 2006. Mitä murteita suomessa onkaan? Murresanaston kvantitatiivista analyysia. *Virittäjä*, 37:245–255.

M. Palander, L. L. Opas-Hänninen, and F. Tweedie. 2003. Neighbours or enemies? competing variants causing differences in transitional dialects. *Computers and the Humanities*, 37:359—372.

I. Savijärvi and E. Yli-Luukko. 1994. *Jämsän äijän murrekirja*. Suomalaisen Kirjallisuuden Seuran toimituksia 618. Suomalaisen Kirjallisuuden Seura.

Tuomi, Tuomo (toim.) 1989. *Suomen murteiden sanakirja. Johdanto*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 36. Kotimaisten kielten tutkimuskeskus.