Multivariate Analysis of Finnish Dialect Data—An Overview of Lexical Variation

Saara Hyvönen
Department of Computer Science, Helsinki Institute for Information Technology, University of Helsinki, P.O. Box 68, FI–00014, Finland

Antti Leino
Department of Computer Science, University of Helsinki, Research Institute for the Languages of Finland

Marko Salmenkivi
Research Institute for the Languages of Finland

Abstract

During the process of writing a comprehensive dictionary of Finnish dialects, a large set of maps describing the regional distribution of the dialect words have been compiled in electronic form. In this article, we set out to analyse this corpus of data in order to gain new insight on the variation of Finnish dialects. We use a wide range of multivariate data analysis methods, including principal components analysis, independent components analysis, clustering, and multi-dimensional scaling. We explain how to preprocess the data to overcome the problem of uneven sampling caused by the way the data has been collected. We discuss the results obtained by these methods and compare them to the traditional view of Finnish dialect groups.

Correspondence:
Saara Hyvönen, Department of Computer Science, Helsinki Institute for Information Technology, University of Helsinki, P.O. Box 68, FI–00014, Finland.
E-mail:
saara.hyvonen@cs.helsinki.fi

1 Introduction

In this article we present computational methods for analysing the geographic variation of linguistic phenomena. The methods outlined here have been applied to the study of lexical variation in the Finnish dialects (Leino *et al.*, 2006); a selection of the results of that study is presented in this article as well.

Our methods are closely related to dialectometry, at least if the term is understood loosely to mean computational dialectology. A computational approach differs from traditional dialectology in that the differences between dialects are studied strictly quantitatively, making it possible to evaluate preconceived notions as to which differences are the 'important' ones.

In recent dialectometric studies, the most common way of measuring the difference between two dialects seems to have been one based on the Levenshtein distance (Levenshtein, 1966). In short, this means recording the same text as spoken by speakers of the different dialects, transliterating these texts and then counting for each pair the number of 1-character differences between the texts. This method has been successfully applied at least to

Literary and Linguistic Computing, Vol. 22, No. 3, 2007.

doi:10.1093/llc/fqm009 Advance Access Published on 2 May 2007

Dutch (Nerbonne and Heeringa, 2001; Nerbonne, 2003) and Norwegian (Gooskens and Heeringa, 2004) dialects. However, such a distance metric cannot be used to analyse a preexisting set of feature distributions.

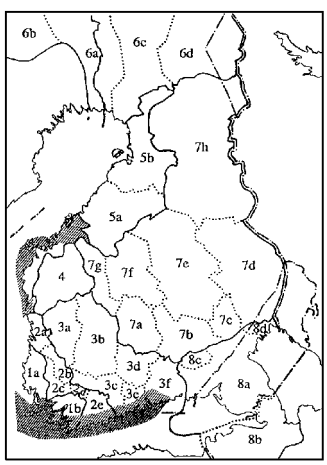
Our case resembles some earlier dialectometric studies of Finnish. Palander *et al.* (2003) studied a transitional zone in the Savonian dialects based on the variation of ten features in the speech of several informants; they applied hierarchical clustering to a matrix that described the frequencies of each of these variants in the speech of their informants. On a geographically wider scale, Wiik (2004) started with a dialect atlas (Kettunen, 1940) and counted the number of isoglosses at each municipality border. The same dialect atlas has also been studied by Embleton and Wheeler (1997, 2000). In their work they describe the process of transforming the dialect atlas into a machine-readable form and their intention of analysing this atlas using multidimensional scaling, but the results of any such analyses have not been widely published.

An important feature of dialect variation is multidimensionality: two dialects can be very close to each other in some respects, yet quite different in others. For example, some of the Finnish dialects along the southwestern coast share a small but significant number of dialect words with Northern dialects along the Swedish border, due to the influence of the Swedish language, although the dialects in other aspects differ strongly. Therefore, multivariate approaches to analysing dialects are needed.

2 Background

According to the traditional view, Finnish can be divided roughly into two dialect groups and these in turn into eight main dialects, as shown in Fig. 1. This division, based mostly on the distribution of the phonological and morphological features that are described by Kettunen (1940), has remained relatively unchanged for the last four decades; the two-way main division is much older, dating back at least to Porthan (1870 [1801]) and arguably even Vhael (1733).

While the division outlined in Fig. 1 is the most common, there are some variations. The second group is a couple of decades more recent than the rest: Rapola (1961) does not yet list it as a separate dialect, but Itkonen (1965) does. More recently, Mielikäinen (1991), joins numbers 4 and 5 into a single group of Ostrobothnian dialects. Even the main division into Eastern and Western dialects has faced some criticism: Paunonen (1991) suggests a three-way split into Western (dialects 1–4), Northern (5–6) and Eastern (7–8) dialect groups, and this was to some extent foreshadowed by Warelius (1848).



Western dialects

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

Eastern dialects

7. Savonian dialects
8. Southeastern dialects

Fig. 1 Finnish dialects (Savijärvi and Yli-Luukko, 1994)

Presenting dialectal variation in this manner is not without its problems, as Heeringa and Nerbonne (2002) point out. This is partly due to the fact that such variation is gradual to some extent. Another reason is that even if one could assume that the distribution of each feature was clearly defined, the isoglosses often do not agree with each other. The traditional way of dealing with this, as at least Bloomfield (1933, 19.8–19.9) already proposes, is to put more weight on 'important' isoglosses and bundles of several isoglosses.

In Finnish dialectology, the traditional yardstick for determining the importance of isoglosses, and the quality of the resulting dialect map, was historical: the ultimate goal was considered to be a map, and an accompanying family tree, showing the relationships of the dialects from the point of view of language history. This gave a rather prominent role to historical phonology on one hand and settlement history on the other. Although dialectology has in the past few decades abandoned much of this philosophical background, it is clear that the resulting research was largely of high quality. Nevertheless, it seems—as Paunonen (1991) points out—that a synchronical study of the dialects should be based on a systematical typological study of the actual variation, not just a collection of features that are deemed important.

One of the main goals of this study is to be such a systematic study of the variation apparent in the lexical data, and in this sense to be an independent evaluation of the established views on Finnish dialects. However, some care should be taken when drawing conclusions, as the earlier studies were mostly based on phonological and morphological variation. Our analysis of lexical variation should not be expected to yield exactly the same results.

2.1 Description of data

Our data consists of the regional distribution of 5,600 dialect words; considering that many of these words have more than one distinct meaning with different geographic distributions, the data amounts to 9,000 distribution maps. These distribution maps have been compiled in the process of writing a comprehensive dictionary of Finnish dialects (Tuomi, 1989), expected to be finished in the 2030s, and they are based on the Lexical Archives of the Finnish Dialects kept at the Research Institute for the Languages of Finland. The original data was mostly collected in the early to mid-20th century.

The first proposals for the dialect dictionary were made in 1868, with a real start in 1896; the first several decades of the project have been outlined by Strandberg (2004). In view of this long history, it is not surprising that the methods for collecting the data have also undergone several extensive changes. Roughly half of the material has been collected by linguistically trained field workers and the other half by volunteers with a widely varying training.

Even with the trained field workers one should remember that the words were collected during roughly three quarters of a century. During that time the methods have varied: at the start, the field workers used questionnaires, while later on the material was collected from free-form interviews. The geographical scope was somewhat limited as well: in the 1920s it became apparent that it was not possible to collect a comprehensive corpus from each of the c. 400 municipalities. Instead, the country was then divided into twenty-three districts, each of which was intended to cover one dialectal region, and the goal was to get a comprehensive collection from at least one municipality in each of these regions (Tuomi, 1989, pp. 18–19).

The result is that the overall collection is far from uniform. Roughly 15% of the municipalities have been systematically surveyed, but even here the number of words from a municipality varies roughly in the range of 10,000–100,000, with 10,000–50,000 having been collected by trained linguists. The collections from the rest of the municipalities are often much smaller.

For the purpose of compiling the dictionary the data is generally quite good: the quality of the field workers appears to have been sufficiently high, and the volunteers were provided with training as well. The spotty coverage is sufficient for the purposes of the dictionary as well. However, it has proved to be one of the main issues in the current study, and one of our major goals is to deal with this widely varying coverage. Another secondary problem is that since we are working with maps created during the

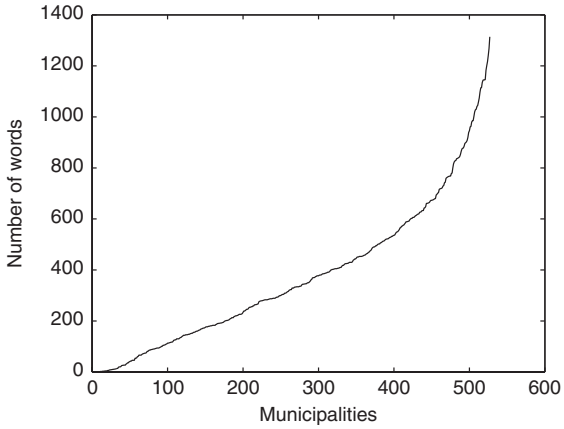


Fig. 2 The number of words per municipality ordered increasingly. In our data, there are 91 municipalities with less than 100 words and 22 municipalities with over 1,000 words

editing process of the dictionary, the data available to us only covers certain sections of the alphabet.

Our data consists of a small portion of the whole collection, namely the words for which a distribution map has been compiled at the time of this work. Maps have not been compiled for words which are either very common or very rare. Figure 2 shows the number of words per municipality in our data. About 25% of the municipalities have less than 150 words, and an equal 25% have more than 520 words.

We have gathered the data into a word-by-municipality-matrix. A small sample of our data is presented in Table 1. If a word has been collected from a municipality, the corresponding element of the matrix is one, otherwise it is zero. Each row corresponds to a word, and gives the regional distribution of this word, that is, which municipalities the word occurs in. Each column tells which words occur in the corresponding municipality. Note that this differs from the format often used by dialectologists, where a row would be a multivalued categorical variable giving the different words used to describe a single concept. Instead our presentation resembles the vector space model approach (Salton and McGill, 1983; Berry *et al.*, 1999) used in analysing text documents: instead of a term-by-document-matrix, we have a word-by-municipality-matrix. Our whole data consists of around 500 municipalities and 9,000 words. About 5% of the entries are ones, the rest are zeros. So our dataset is large, sparse (only a small portion of the elements are non-zero) and binary (all elements are either 0 or 1).

Table 1 A sample of our data. Vihti and Aura are municpalities in western Finland, Kitee and Juva are located in eastern Finland. The value 1 in the table indicates that the word occurs in the corresponding municipality

	Vihti	Aura	Kitee	Juva
aprakka	0	0	1	1
epatto	0	0	1	1
filunki	1	1	0	0
haalakka	0	0	1	1
hampuusi	1	1	0	0
kräki	0	1	0	0

3 Methods

We have analysed our dialect data using multivariate methods for data analysis. We describe here a few such methods: principal components analysis (PCA), independent component analysis (ICA), multidimensional scaling (MDS), and clustering. These are all standard methods in the field of data analysis, and can be found in a number of text books (e.g. Hand *et al.*, 2001; Hastie *et al.*, 2001; Hyvärinen *et al.*, 2001). However, while some of these methods have been successfully applied to analyse dialect data, e.g. by Shakelton (2005), applying them to this kind of data is not always straightforward; therefore it is useful to discuss them here in some detail. Furthermore, when applied to the raw data, multidimensional scaling and clustering do not perform very well. This is due to the uneven sampling rate of the data: as described in Section 2.1, the number of words collected from each municipality varies a lot. This problem can be overcomed by preprocessing the data in an appropriate manner, as described in Section 3.3.

Table 2 The principal components computed from the sample data presented in Table 1. The first principal component separates east and west. The second principal component captures the difference between the two western municipalities

Municipality	1st principal component	2nd principal component
Vihti	−0.40	−0.91
Aura	−0.53	0.23
Kitee	0.53	−0.23
Juva	0.53	−0.23

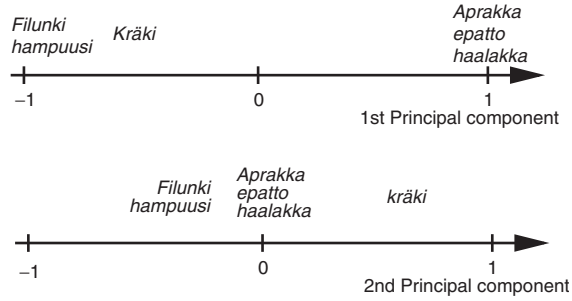


Fig. 3 The position of the words projected onto the 1st and 2nd principal components. Again the first principal component separates east and west, while the second principal component captures the difference between the two western municipalities

3.1 Principal components analysis

The aim of PCA (Hotelling, 1933) is to capture the intrinsic variability in the data. We seek to find a combination of features, in our case of municipalities, which explains the variance in the data as well as possible. This is the first principal component. We then proceed by repeatedly looking for a combination of features which is uncorrelated with all of the previous principal components and which explains as much of the remaining variance in the data as possible.

In our case each principal component is a vector, the elements of which correspond to the municipalities. The weight of each element is computed so that a maximal amount of the unexplained variation in the data is captured by the principal component. Thus, each element describes the role of the municipality in the corresponding principal component, and the principal component itself describes the variation in the data.

For example, if we apply PCA to the sample data given in Table 1, we get the principal components given in Table 2. The first principal component, explaining 75% of the variation in the data, separates the western municipalities (negative values) from the eastern ones (positive values). The second principal component, explaining the remaining 25% of the variation in the data, captures the difference between the two western municipalities Aura and Vihti.

The first principal component is the (linear) combination of the original variables, here municipalities, which captures the maximum amount of the variance in the data. For our example case this is illustrated in Fig. 3. Most of the variance in the data is due to the east-west division, so the first principal component separates the eastern words from the western ones. The second principal component tries to explain as much of the remaining variance as possible. In our example, only the difference between the two western municipalities remains to be explained. Thus the second principal component separates the word *kräki*, which occurs only in Aura, from the other two western words.

In mathematical terms, computing the principal components involves computing the singular value decomposition (SVD) of the centred data matrix. Our data matrix is of the form described in Table 1, with rows representing the data points (here words) and the columns representing variables (here municipalities). We first centre the data by subtracting the estimated mean from each column. Then we compute the singular value decomposition of the centred data matrix. The right singular vectors are the principal components. The singular values give the variance captured by the corresponding principal components. In terms of preserving the variance, the best description of the data that can be given using, say, three variables is given by the three first principal components. Similarly, if we want to describe our sample data in Table 1 using only one variable, the best we can do is to project the data onto the first principal component, as was done in Fig. 3. The PCA algorithm is incorporated in most statistical software packages.

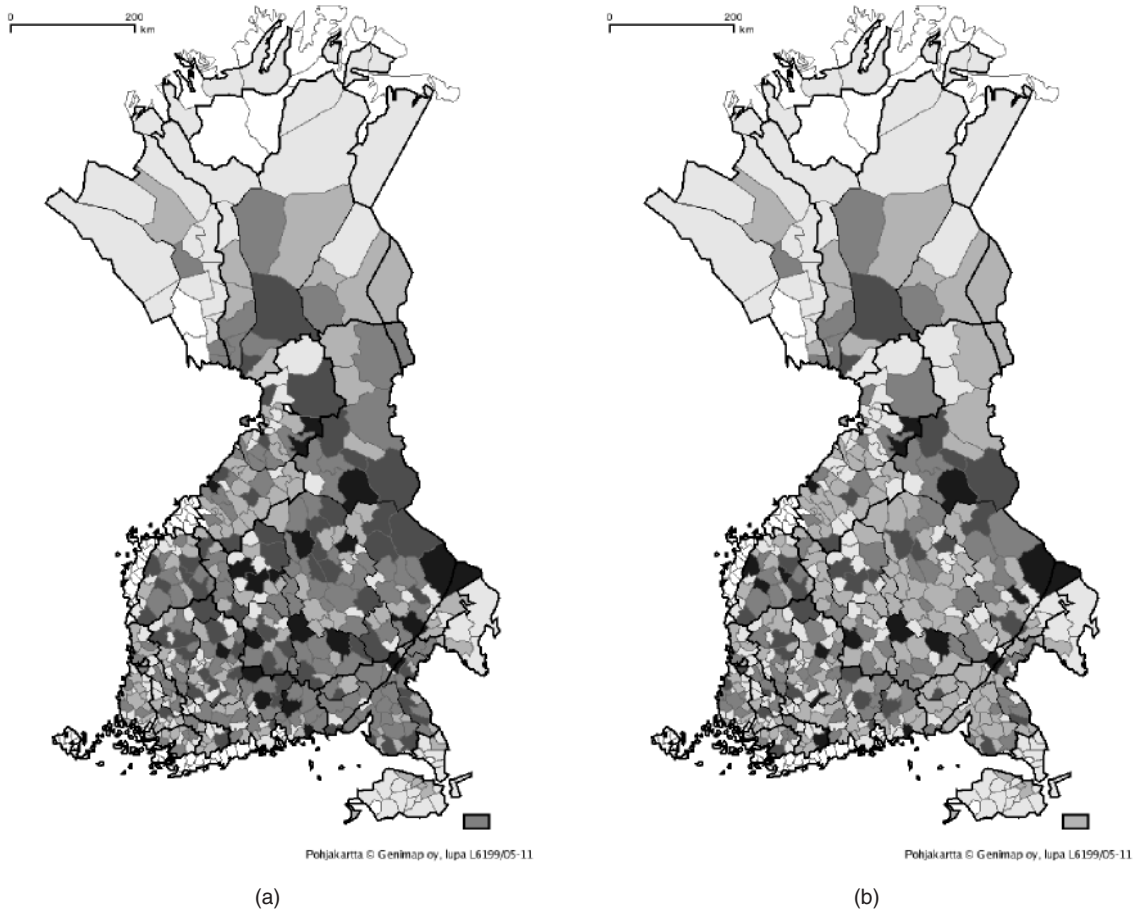


Fig. 4 (a) First principal component. (b) Number of words

We can visualize each principal component by plotting a map where each municipality is coloured according to the value the corresponding element of the principal component has. The municipalities with the most extreme values on each principal component are coloured black or white, and the municipalities with less weight in the component are plotted in shades of grey. In our example case presented above, the first principal component would be visualized by a map such that Aura would be white, Juva and Kitee would be black, and Vihti would be light grey. The map corresponding to the second principal component would have Aura coloured black and Vihti white, and Juva and Kitee would be grey. It is not significant which extreme is black and which is white, the roles of these could as well be reversed. The interesting thing is comparing the relative geographical position of the extremes: this shows which municipalities or which regions are important in explaining the variance in the data.

The principal components for our data are presented in Figs. 4–6. The first principal component describes essentially the number of words collected from each municipality, as can be seen by comparing Figs. 4a and b. This is the most important factor in explaining the variance in the data—but not really what we are likely to be interested in. After this we see more interesting things: the second principal component tells us about the division between east and west in Finnish dialects, then in the third the north-south division is apparent.

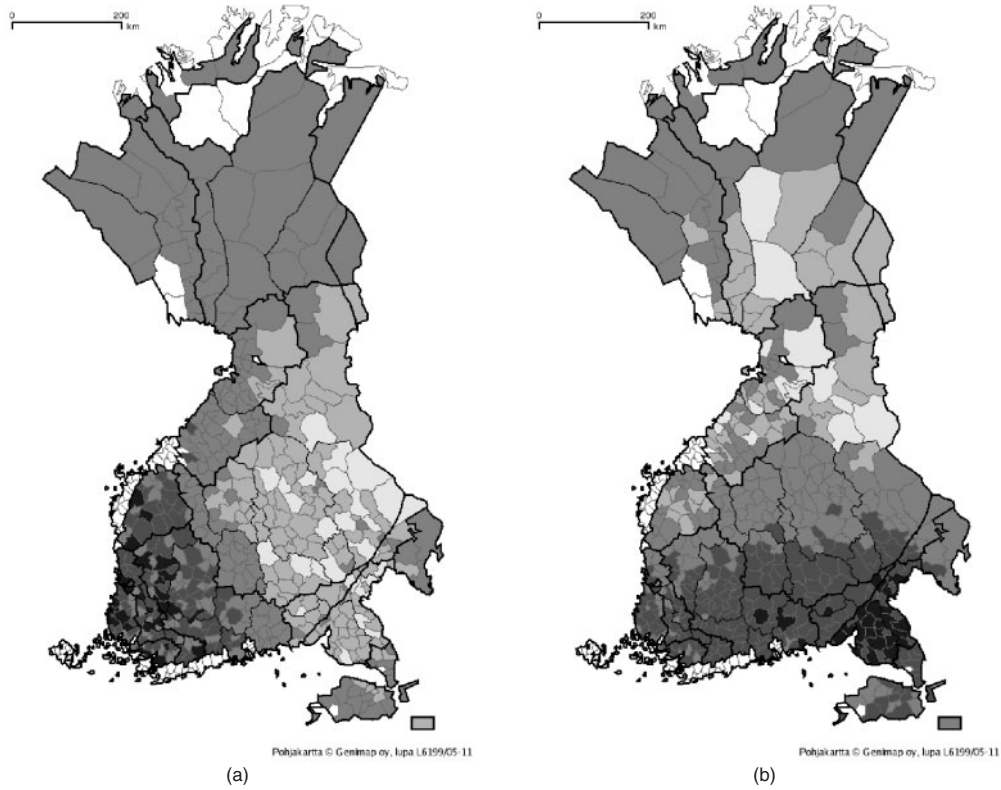


Fig. 5 (a) Second and (b) third principal components

The linguistic interpretation of the principal components is discussed in more detail in Section 4.2.

3.2 Independent component analysis

While PCA aims to find a low dimensional representation of the whole data set in such a way that a maximal amount of the variance in the data is preserved, ICA has a different goal. ICA seeks to model the data as linear mixtures of some underlying factors or components (Hyvärinen 1999; Hyvärinen *et al.*, 2001). In this context these factors could be thought to represent different dialects, and our goal would be to represent the data by summing up the contributions of separate dialects. ICA is closely related to a statistical method called factor analysis (FA), which has been used in analysing dialect data (Nerbonne, 2006); in fact, ICA can be considered as non-gaussian factor analysis. What distinguishes ICA from FA and other similar methods is that it looks for *statistically independent* components in *non-gaussian* data. For more details on the relationship between ICA, PCA, and FA, see e.g. (Hyvärinen *et al.*, 2001).

Real data often does not follow a gaussian distribution, and in such cases statistical independence is a stronger requirement than uncorrelatedness, required by PCA and FA. In particular, dialect data had the property that there are groups of words typical to one dialect which only appear in one region: it is this kind of non-gaussian on/off behaviour of the data that the independent components will capture. In our case each element of an independent component again corresponds to a municipality. These can be visualized as maps in the same way as the principal components.

PCA is used as a preprocessing step in ICA. We discard the first principal component which only tells us of the sampling rate and then include a number of principal components equal to the number of independent components we wish

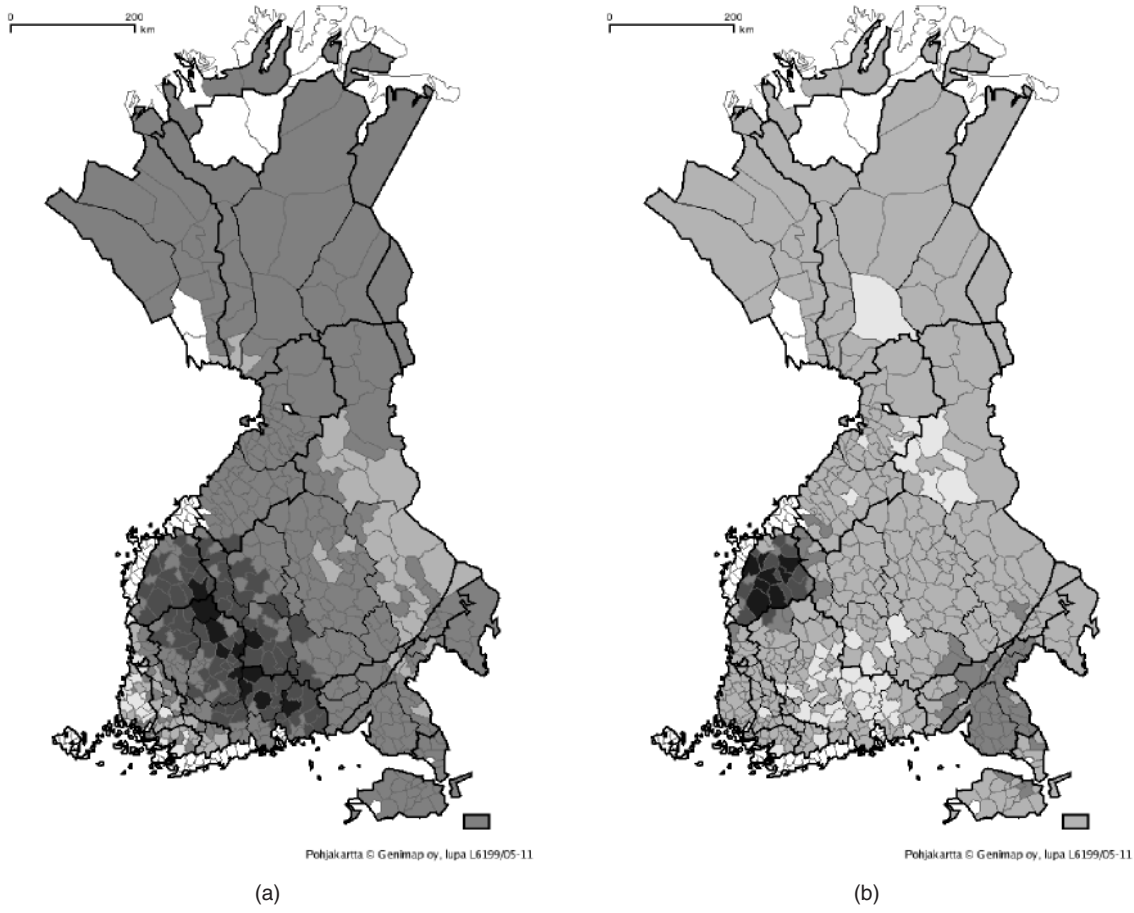


Fig. 6 (a) Fourth and (b) fifth principal components

to discover. Compared to principal components, the independent components have a more local structure. For example, the 6 independent components presented in Figs. 7–9 correspond fairly nicely to a separate dialect region each. Compare these to the principal components in Figs. 4–6 that present a more global structure, e.g. east versus west in PC 2 instead of south-west versus the rest of Finland in IC 2.

Notice that unlike PCA, ICA does not rank the components, but they all are of equal importance. Furthermore, the number of independent components desired usually needs to be specified in advance. It is difficult to give theoretical guidelines for choosing the number of components (Hyvärinen *et al.*, 2001, pp. 269–271), so this is often done by trial and error. In our application, the larger the number of components, the more localized they tend to be. Not all choices yield equally meaningful results.

3.3 Preprocessing data via singular value decomposition

Many methods useful for visualizing dialect variation require us to be able to determine the similarity or difference of two municipalities. Such methods include clustering and multidimensional scaling, both of which are discussed later in this article. But in the case of our data determining the similarity of two municipalities is non-trivial. For example, assume one municipality has been

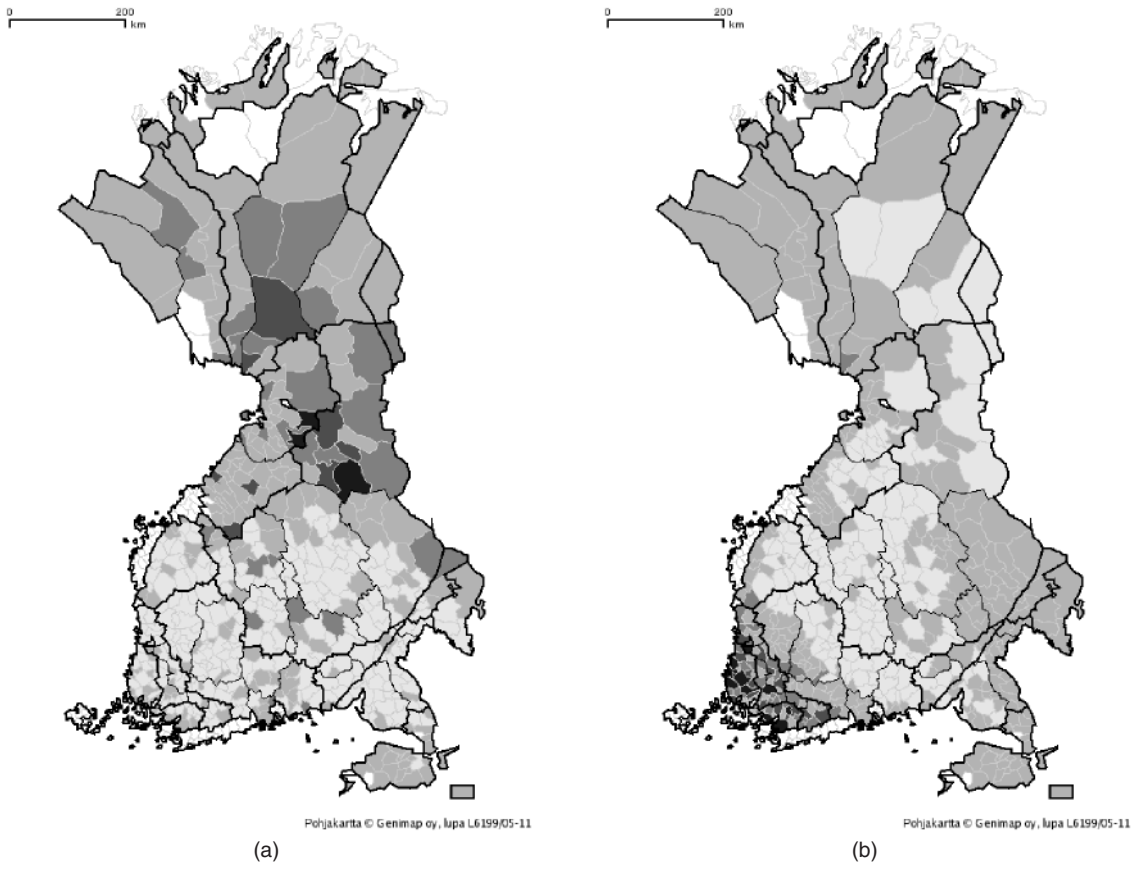


Fig. 7 (a) First and (b) Second independent components

well studied, while its neighbour has been very poorly studied, so that in the latter case we know only a few words present in that municipality. The two municipalities will look quite different just because of the uneven sampling. Leaving out the municipalities with very few words does not help: there will always be a new category of 'municipalities with few words', our definition of few just goes up (Fig. 2). One way to overcome this problem is to use the cosine distance measure instead of the more commonly used Euclidean distance. The cosine measure uses the angle between vectors to judge similarity; the lengths of the vectors, which roughly speaking are related to how well the municipalities (which those vectors represent) have been sampled, are not important. But even this will not work if we have two municipalities from the same dialect region, but just happen to sample completely different words from them. Something akin to this might indeed happen in our data, due to the colourful history of collecting the data described in Section 2.1.

As an example consider the data sample in Table 3, which is an extension of Table 1. There are three western municipalities with unevenly sampled data. The two eastern municipalities are well sampled. If we compute the cosine similarities from this data directly, we get the results prestented in Table 4. The uneven sampling clearly disturbs the results, as we would like to see the western municipalities as similar to each other. We need to preprocess our data to obtain more meaningful results.

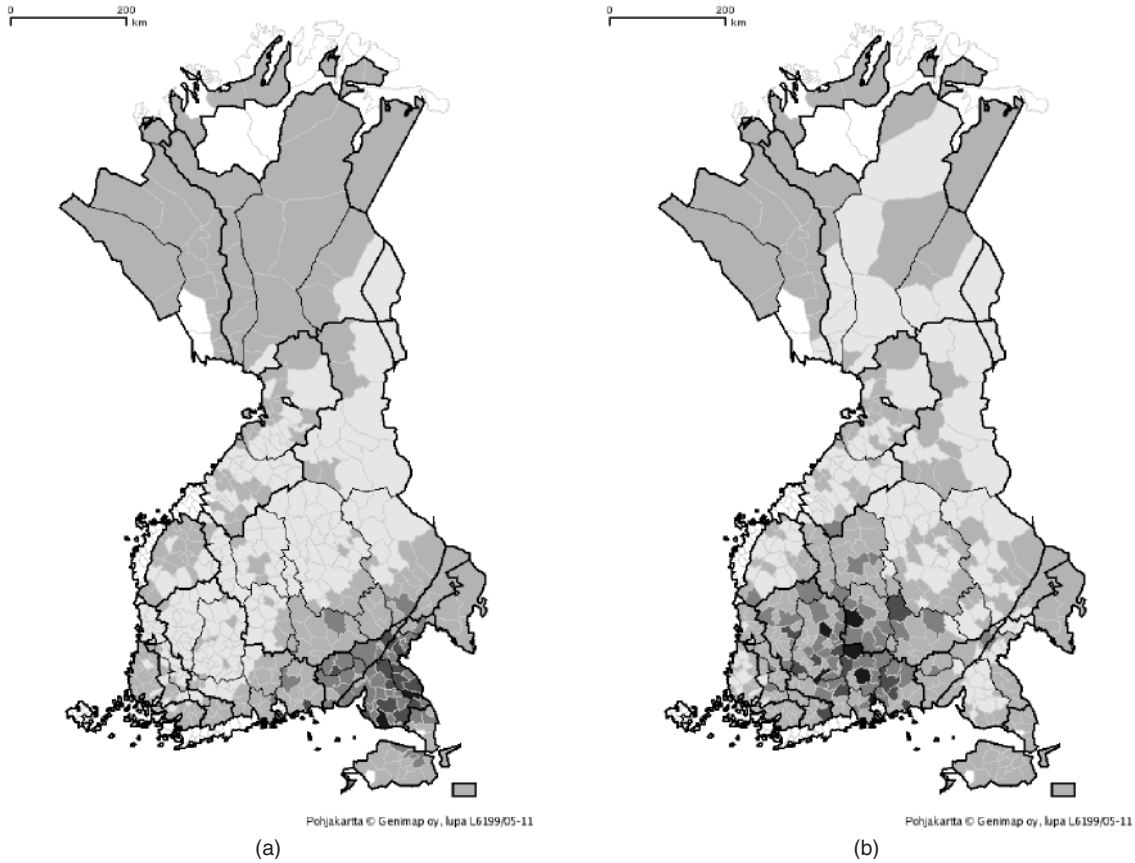


Fig. 8 (a) Third and (b) fourth independent components

We already noted in Section 2.1 that our problem resembles that of analysing large sets of text documents, which is a widely studied problem (Berry *et al.*, 1999). A common approach in these applications is to compute the truncated SVD and use this instead of the original matrix when analysing the data. This means that instead of describing each municipality by using all the dialect words we 'summarize' the words into so-called singular vectors and express the data in terms of a few of them. This approach is also known as latent semantic indexing (LSI) (Deerwester *et al.*, 1990) or latent semantic analysis (LSA) (Landauer *et al.*, 1998). This approach has been used with varying success in e.g. language processing applications (Schone and Jurafsky, 2000; Baldwin *et al.*, 2003). LSI is based on the assumption that there is some underlying, latent structure in the data, such as dialect variation in our case, and the truncated SVD is used to estimate this structure. LSI is used both to speed up computations (by reducing the size of the data set) and to improve the quality of the data (by removing extraneous information or noise in the data).

Let us return to the example in Table 3. Instead of using the original data to compute the similarities between municipalities let us use LSI and present the data using two singular vectors, and then compute the similarities. The results are presented in Table 5. This time all western municipalities are similar to each other, as are the eastern ones. This is because though Kaarina is badly sampled, there is a municipality in the same dialect region, namely Aura, which is well sampled. Vihti and Kaarina

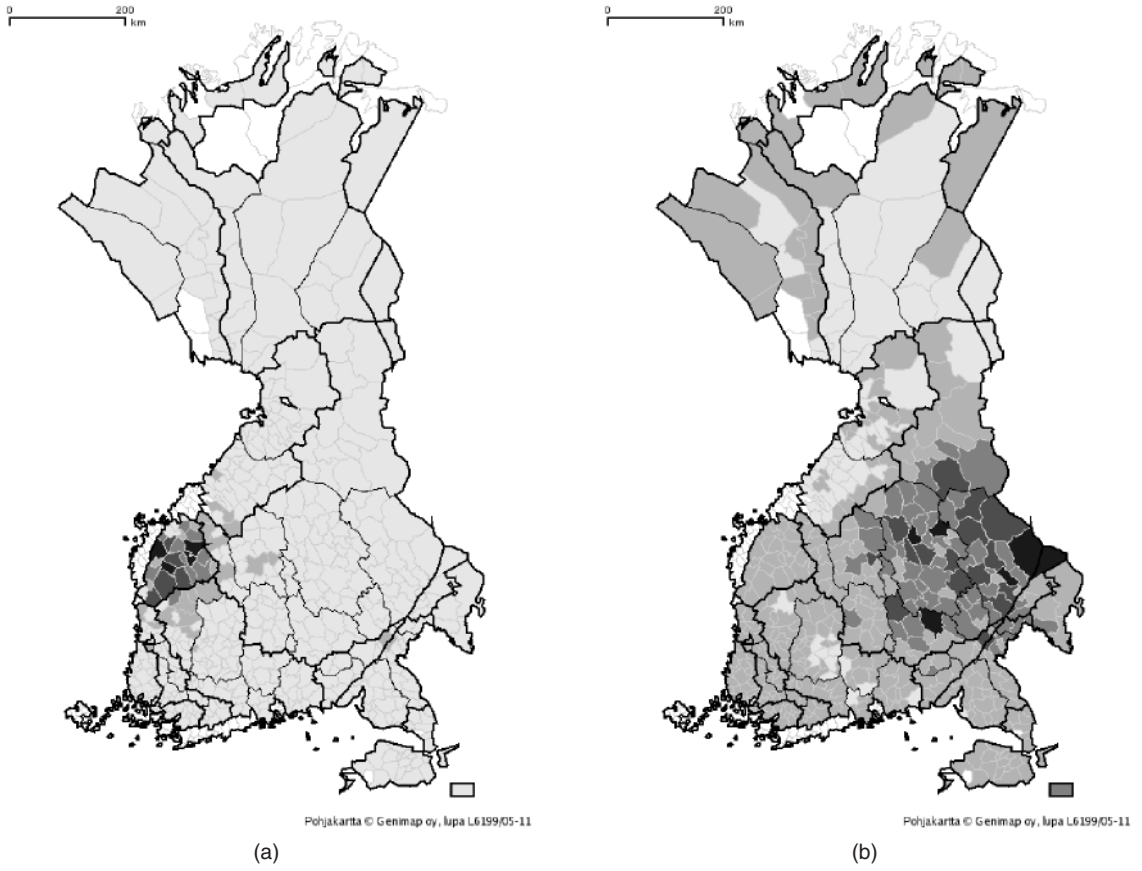


Fig. 9 (a) Fifth and (b) sixth independent components

Table 3 An extension of the data sample given in Table 1. Vihti, Aura, and Kaarina are municipalities in western Finland, Kitee and Juva are located in eastern Finland. The value 1 in the table indicates that the word occurs in the corresponding municipality

	Vihti	Aura	Kitee	Juva	Kaarina
aprakka	0	0	1	1	0
epatto	0	0	1	1	0
filunki	1	1	0	0	0
haalakka	0	0	1	1	0
hampuusi	1	1	0	0	0
kräki	0	1	0	0	1

Table 4 The cosine similarities of the municipalities. Though Vihti, Aura, and Kaarina are all western municipalities, their pairwise similarities vary, due to the uneven number of words sampled from each municipality. The eastern municipalities are both well sampled, so their similarity is high

	Vihti	Aura	Kitee	Juva	Kaarina
Vihti	1	0.82	0	0	0
Aura	0.82	1	0	0	0.58
Kitee	0	0	1	1	0
Juva	0	0	1	1	0
Kaarina	0	0.58	0	0	1

appear similar because they are both similar to the same, well-sampled municipality, Aura. With Aura missing, using LSI would not help. Note that this example was selected to illustrate the effect of uneven sampling: in reality a single word will of course not have such a great impact on the results.

But why do we use the singular vectors to summarize the data instead of e.g. principal components?

Table 5 The cosine similarities computed in the SVD basis using two singular vectors. The effect of the sampling rate disappears

	Vihti	Aura	Kitee	Juva	Kaarina
Vihti	1	1	0	0	1
Aura	1	1	0	0	1
Kitee	0	0	1	1	0
Juva	0	0	1	1	0
Kaarina	1	1	0	0	1

There are several reasons. First of all, the truncated SVD gives a mathematically well justified representation of the data; more precisely, it gives an optimal low-rank approximation of the original matrix in the sense that it minimizes the norm of the error matrix. But here, in fact, the principal components and the singular vectors are very close to each other. Remember that computing the PCA involved computing the SVD of the centred data matrix. Centreing means subtracting the mean from each column. Here most entries of the matrix are zero, so the mean of each column is close to zero, and centreing does not change much. But it does change something: centreing the data can move points close to each other to different sides of the origin. Then the cosine distance will become large, even if the points in the original space are close to each other.

Though our problem resembles that of analysing large sets of text documents, differences still remain. We only have the information of whether a word occurs in a municipality or not, and even that information suffers from the uneven sampling rate. In the case of text documents the precise number of times each word appears in a document can be counted. Therefore it is possible to devise schemes, which give the terms different weights depending on how common or rare they are (Salton and Buckley, 1988); such techniques are not suitable for our data.

Like the first principal component (see Section 3.1), the first singular vector correlates strongly (98%) with the number of words occurring in each municipality. Thus it tells about the sampling rate, and little else. Hence, it makes sense to drop this singular vector and thereby eliminate, or at least reduce, the effect of the sampling rate. So we project our data onto the space spanned by the first singular vectors starting from the second one. In these analyses, we have used the singular vectors 2–30.

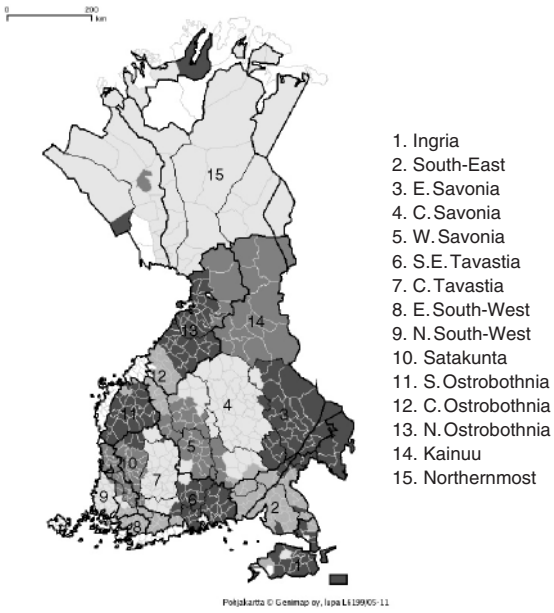


Fig. 10 Partitioning into fifteen clusters

It is worth mentioning that even this approach does not rid us completely of the problem caused by uneven sampling. For example, if Aura was missing from Table 3, Vihti and Kaarina, the two other western municipalities, would not be judged similar. In our data set the undersampled municipalities in south-western Finland do get taken care of, because there are lots of well-sampled municipalities there. But some regions outside the pre-World War II borders of Finland, e.g. Ingria, are completely undersampled: there are no well-sampled municipalities. Therefore, these regions cannot be well represented using the singular vectors either. This explains the outliers in Fig. 10.

3.4 Clustering

Clustering methods aim to divide the data points into a predetermined number of clusters in such a way that objects in the same cluster are similar to each other, while objects in different clusters are dissimilar. Consider again our sample data

in Table 1. We see that the two eastern municipalities Kitee and Juva have identical word profiles, while those of the western Aura and Vihti resemble each other. Thus, clustering the data into two groups would produce an eastern and a western cluster. A division into three clusters will separate Aura and Vihti into clusters of their own.

3.4.1 *K-means clustering*

A widely used clustering method is the K-means algorithm (MacQueen, 1967). In the basic version one starts by randomly picking K cluster centres. One then repeatedly assigns to each cluster all points closest to the appropriate cluster centre, and recomputes the new cluster centre as the mean of all points in that cluster. This is done until no changes in the centres occur. In our application the points are the word occurrence vectors of the municipalities. The cluster centre is obtained by averaging over the word occurrence vectors of the municipalities belonging to that cluster. It can be thought of as a typical representative of the cluster.

However, straightforward application of K-means to our data is problematic for two reasons. First of all, the data set is fairly large, and so the algorithm is slow. More importantly, the uneven sampling leads to a situation in which municipalities with very few words tend to end up in the same cluster. Both of these problems are solved by the preprocessing scheme described in Section 3.3 and by adopting the cosine distance measure. Ergo, we process our data set by doing a truncated SVD to remove noise while still retaining most of the variation in the data. Furthermore, we discard the first singular vector and with it the variation which is due to the uneven sampling rate. Our preprocessed data still consists of one vector per municipality, but now the elements of the vectors no longer correspond directly to the dialect words, but to the singular vectors, which in a sense summarize the information in the data. We can apply data analysis methods, such as clustering, to this new data set. We have used K-means clustering with the cosine measure on this data.

In many cases one is interested in finding the correct number of clusters, and many strategies for choosing this are available, see e.g. (Theodorids and Koutroumbas, 2003). For example, the Davies–Bouldin index (Davies and Bouldin, 1979) is a function of the ratio of the sum of within-cluster variation to between-cluster separation, and therefore favours compact and well-separated clusters. Dialect data, however, tends to have a hierarchical structure, so the question of the 'correct' number of clusters depends on the amount of resolution we are interested in. With just two clusters we obtain the traditional division to Eastern and Western dialect regions. As we increase the number of clusters, we get a more refined partitioning as smaller dialect regions form separate clusters. By comparing how the boundaries of these regions change as the number of clusters is increased we can evaluate the robustness of each boundary.

3.4.2 *Hierarchical clustering*

Another commonly used clustering method is hierarchical clustering (Jain and Dubes, 1988). There are two basic approaches to hierarchical clustering. Agglomerative clustering starts with each point forming a separate cluster. After this the two clusters closest to each other are repeatedly merged together until only one cluster, including all the points, exists. This form of hierarchical clustering has been used, e.g. to cluster Dutch dialects (Heeringa and Nerbonne, 2002), Irish Gaelic (Kessler, 1995), and Tyneside English (Moisl and Jones, 2005). There are several alternative methods to decide when to merge two clusters, among them the average linkage method and Ward's method (Kaufman and Rousseeuw, 1990).

An alternative approach is provided by divisive clustering (Kaufman and Rousseeuw, 1990, pp. 253–79), which proceeds in an inverse order. At each step, a divisive method splits up a cluster into two smaller ones, until all clusters contain only a single element.

An advantage of the hierarchical approach is that it not only gives a partitioning of the municipalities into dialect regions, but also produces a 'family tree' of the dialects. One might assume that hierarchical clustering works well with dialect data, due to its inherently hierarchical structure. Our results, however, do not confirm this. None of the approaches work particularly well. This is mainly

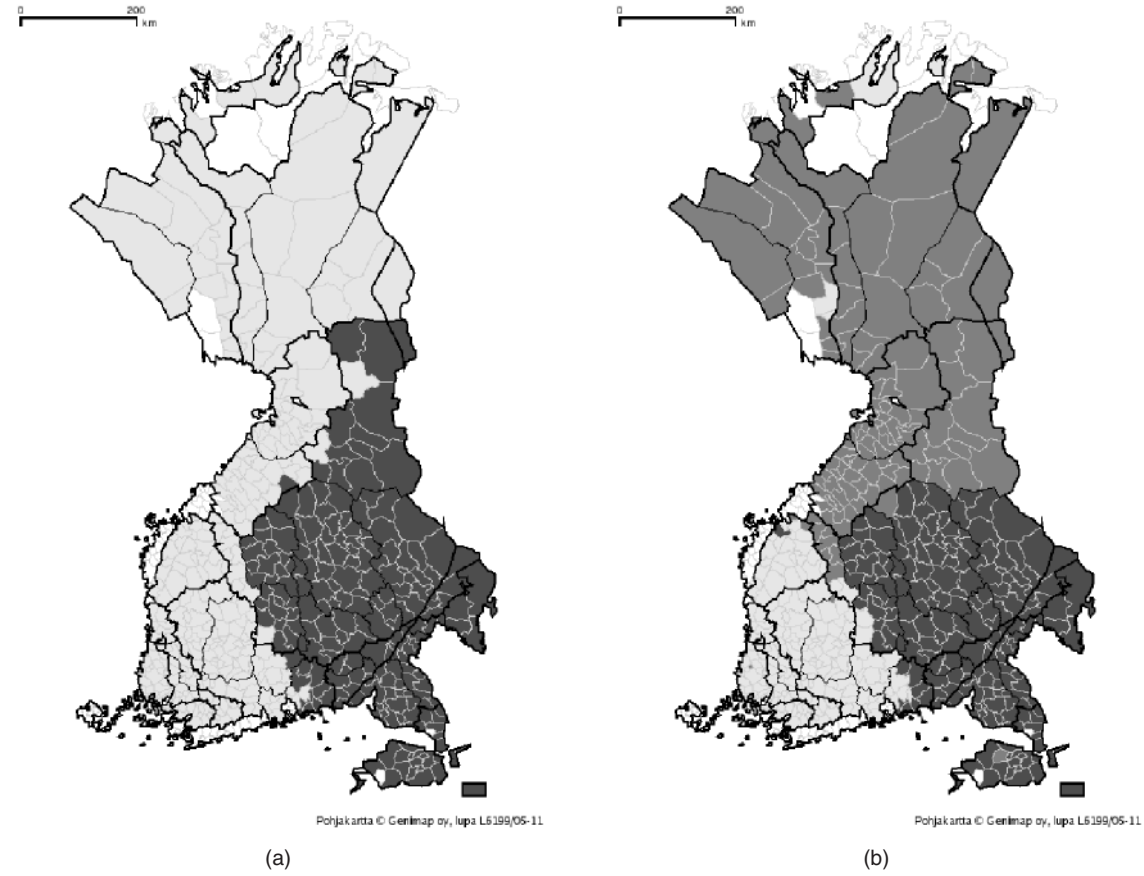


Fig. 11 Partitioning into (a) two and (b) three clusters

due to the fact that once a decision is made during the clustering process, it cannot be reversed. For instance, the divisive method makes an initial division into two regions similar to that given by the K-means method (see Fig. 11a), and is unable later to handle the Northern dialect region properly, as the division into east and west cannot be reversed in that area. Furthermore, hierarchical methods also seem slightly more sensitive to outliers than K-means.

However, we do show how hierarchical clustering can be applied to the dialect regions. Each dialect region is represented by the cluster centre obtained using K-means clustering, and using hierarchical clustering on these and examining the resulting dendrogram gives us information about how the dialect regions relate to each other. We have used Ward's method in the case of agglomerative clustering. Note that here we have used the Euclidean distance between points, as in theory Ward's method requires us to do so[1] (Kaufman and Rousseeuw, 1990, 230–34).

The results for both divisive and agglomerative clustering are presented in Figs. 12 and 13. Note, that in the case of agglomerative clustering once two clusters have been merged, they cannot be separated later. This can have a significant impact on transitional dialects: for example in Fig. 13 Western Savonian and Southeastern Tavastian end up together, which results in the latter being classified as an Eastern dialect. On the other hand, once two regions have been parted in divisive clustering, they cannot be joined together again, so though obviously Western Savonian and Southeastern Tavastian

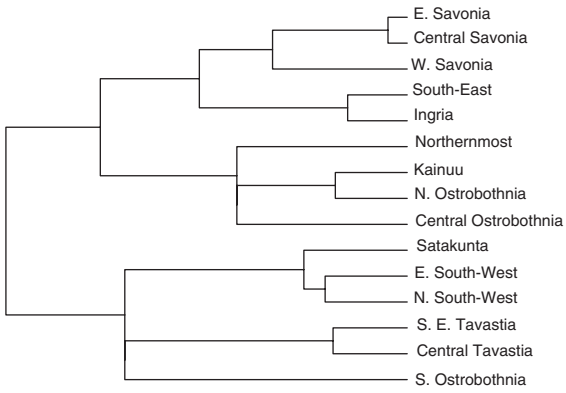


Fig. 12 Hierarchical clustering of the fifteen dialect regions obtained by using K-means clustering on the whole data. Here a divisive strategy was used

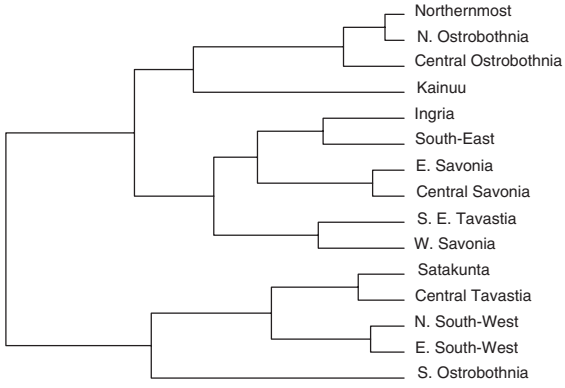


Fig. 13 As in Fig. 12, this time an agglomerative strategy with Ward's method was used. Note that Western Savonian and Southeastern Tavastian end up together

are fairly close to each other, they are separated at the very first stage of divisive clustering, so their similarity is lost in Fig. 12.

3.5 Multidimensional scaling

MDS is a statistical technique (Mardia *et al.*, 1979) that has been used in analysing dialect data (Embleton and Wheeler, 1997; Nerbonne *et al.*, 1999; Heeringa and Nerbonne, 2002). The basic idea (Xu and Wunsch, 2005) lies in fitting original multivariate data into a low-dimensional structure while aiming to maintain the proximity information. MDS uses the pairwise distances between points. Here again it is useful to preprocess our data as described in Section 3.3 and use the cosine distance measure. Again we shall not present the results for the whole data set (which are obtainable from the authors directly) here, but instead demonstrate the use of MDS by plotting onto a plane the fifteen dialect regions given by K-means clustering. This is shown in Fig. 4. We see that the resulting map bears a strong resemblance to the geographical map in Fig. 1. Note in particular that Southeast Tavastia, which when using hierarchical clustering landed in the Eastern dialect group, is here somewhere midway between east and west, slightly closer to Central Tavastia than to West Savonia.

While MDS seeks to place the points on the plane in such a way that distance information is preserved as well as possible, this is still a simplification of the real situation. A slightly more complex view is provided by using e.g. three dimensions instead of only two.

4 Results

Our results agree rather well with the traditional view. Considering how different our approach is, this is independent validation of the results of prior research done with traditional methods. There are, however, some areas where the results disagree, and most of these can be explained either by the different diffusion of lexical versus morphological or phonological phenomena or by the importance placed on settlement history by earlier scholars.

4.1 Dialect areas

In dividing Finnish into dialects, our main disagreement with earlier research is on the topmost level. Instead of dividing Finnish into an Eastern and a Western dialect group we propose three groups: a three-way resulting in Eastern, Western, and Northern dialects describes the data much better. Still, a two-way clustering splits the area roughly along the traditional border, as seen in Fig. 11a, so such a division is not altogether

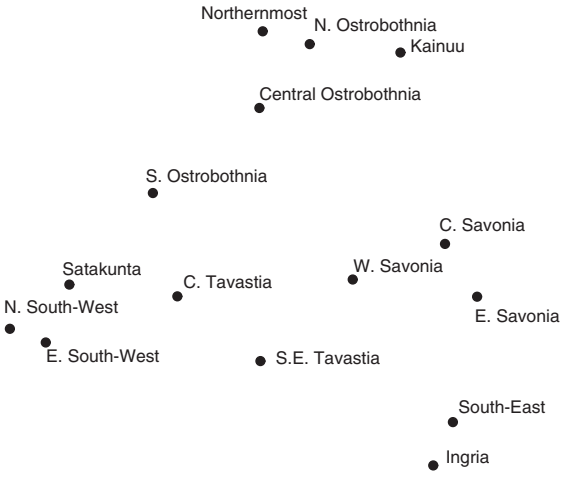


Fig. 14 The fifteen dialect regions from Fig. 12 mapped onto the plane using MDS

Table 6 A summary of the PCA results. The thirty first principal components explain 43% of the variance

PC	Figure	% of variance explained	Description
1	4a	16.2	Number of words
2	5a	5.5	East-west variation
3	5b	2.5	North-south variation
4	6a	2.1	Tavastia
5	6b	1.9	Southern Ostrobothnia

inconceivable, but the borders of the three-way split shown in Fig. 11b prove much more stable when increasing the number of clusters.

The three-way division is also clearly seen in Fig. 14. Here, multidimensional scaling results in a clear separation of the Western, Eastern, and Northern dialect groups, with the transitional dialects between the Eastern and Western appearing within the gap between the two. It is also clear from the figure that there is a clear division between the Southern Ostrobothnian dialect and the rest of the Western group.

While hierarchical clustering creates two-way divisions, the agglomerative dendrogram in Fig. 13 can also be seen as evidence against the traditional view. Here, the first division is between the Western dialects on one side and the Eastern and Northern on the other, but the Northern dialects are the first to split from this cluster. Similarly, the Southern Ostrobothnian are the first to split from the Western cluster, again as one would expect after looking at the MDS results.

The divisive dendrogram in Fig. 12 also makes the first division between the Western dialect group and the other two, and the next one between the Eastern and Northern. The internal structure of the Western and Northern groups comes out rather different, and the transitional Southeastern Tavastian and Western Savonian dialects come out rather different. This, in turn, makes the division of the Western dialect group somewhat messy, as the Tavastian dialects now appear much more as a distinct sub-block.

Related to this change from a two-way to three-way main division, our data suggests a different grouping for the Kainuu dialects, number 7h on the map in Fig. 1. Traditionally these dialects have been considered a part of the Savonian dialects, and a lot of the Savonian features extend to Kainuu. Still, as Wiik (2004, pp. 25–7) shows, almost as many isoglosses separate the area from the more southern Savonian dialects as from the Ostrobothnian and Northernmost dialects. Lexically, however, the dialects are much closer to the Ostrobothnian and they clearly are a part of the Northern dialect group. This can be seen both in the MDS results, Fig. 14, and in hierarchical clustering, Figs. 12 and 13, which all put Kainuu in the Northern dialect group and not the Eastern one.

There are also some differences in how the various transitional dialects are treated. In the Savonian dialects, the Keuruu–Evijärvi area (number 7g in Fig. 1) has been hard to classify all along, and Rapola (1961) already notes that the dialects are only half-Savonian. On the other hand, its border against the Southern Ostrobothnian dialect (number 4) is the sharpest in all Finnish dialects (Wiik, 2004, p. 45). Our analysis puts most of the area into the Central Ostrobothnian cluster and thus into the Northern main group, with just the southernmost part in the Western Savonian dialects, see Fig. 10.

In the Western dialect group we see the disappearance of an entire main dialect. The Mid-southwestern dialects (number 2) do not appear as

a separate group at all: instead, the northern part of the area joins the Western Tavastian dialects and the southern part joins the eastern half of the Southwestern dialects, as seen in Fig. 10. This is, in fact, the same order in which Rapola (1961) discusses these dialects.

All in all, our analysis results in the following main dialects:

Western dialects

(1) Southwestern dialects
(2) Tavastian dialects
(3) Southern Ostrobothnian dialects

Eastern dialects

(4) Savonian dialects
(5) Southeastern dialects

Northern dialects

(6) Ostrobothnian dialects
(7) Northernmost dialects

For the most part, the borders between these dialects match the traditional dialect borders; the differences show in their classification. This is a rather interesting discovery in itself, considering how traditional wisdom tells us that the diffusion patterns of lexical phenomena are often very different from those of morphological/phonological ones.

4.2 Components in the dialectal variation

Although clustering can be used to find dialectal regions, shown both on a map and in the form of a tree, this does not tell us everything there is to know. The variation consists of a series of diffusion patterns, and two dialects that are close to each other in some respects may be quite different in some others. Methods like or PCA/ICA make it possible to summarize the thousands of individual distributions as a few underlying components.

Figures 4–6 show the first five principal components. The amount of variation explained by these is shown in table 6. As noted earlier, the first principal component, shown in Fig. 4a, is very highly correlated with the overall number of words in the municipality, shown in Fig. 4b. It is essentially noise, but it was good to be able to separate it, as the corresponding singular vector could be discarded as a preliminary step before clustering.

The two components shown in Figs. 5a and b correspond to the two main variations in the data. The second principal component corresponds with that between the Eastern and Western dialects. As this has traditionally been considered the main dialect division in Finnish, it would have been quite surprising if it had not turned up as the first linguistically defined component. Similarly, the variation between Northern and Southern dialects is seen in the third principal component. Here the northern extreme is on the region between the Kainuu and Northern Ostrobothnian dialects, possibly because the data from more northern municipalities is much more sparse. The southern end is in the Southeastern dialects on the Carelian isthmus.

The fourth and fifth principal components, shown in Figs. 6a and b, each appear to define a single main dialect. Of these, the fourth appears to indicate a Tavastian, as opposed to Southwestern, influence. The fifth principal component is geographically much more compact. It is almost completely limited to Southern Ostrobothnia, and this is in itself a strong indication of how much this dialect differs from its neighbours: according to Wiik (2004) the northeastern border of this dialect is the sharpest in Finnish, and the southeastern border is not very far behind.

Independent components, on the other hand, organize the data somewhat differently. Figs. 7–9 show the results of an analysis where six components were extracted from the data; it is instructive to compare these to the results of both the clustering and the principal components analysis discussed above.

On the one hand, each of the maps shows a relatively clear darker region, and these correspond well with the primary dialect regions: component 1 is quite close to the Northern main dialect group, component 2 the Southwestern dialects, component 3 the Southeastern dialects, component 4 the Tavastian dialects, component 5 the Southern Ostrobothnian dialects and finally

component 6 the Savonian dialects. These six are also the first six dialect groups one gets from a divisive clustering of the fifteen dialect regions, as seen in Fig. 12.

On the other hand, it is also instructive to compare the independent components to the principal components. Here, principal component 2, in Fig. 5a, can be viewed as the contrast between independent components 2 and 6, in Figs. 7b and 9b. Similarly, it is possible to see in principal component 3, in Fig. 5b, the combination of independent components 1 and 4, in Figs. 7a and 8b. Under this interpretation, principal components 4 and 5, in Figs. 6a and b, correspond to independent components 4 and 5, in Figs. 8b and 9a.

All in all, PCA tends to capture global variation, whereas ICA finds more local phenomena. On the resulting maps the independent components are sharper and geographically more compact than the principal components. Increasing the number of independent components to be found leads to even more compact components, corresponding to smaller dialect regions. For example, most of the components obtained when computing fifteen independent components correspond closely to one of the clusters presented in Fig. 10.

5 Conclusions

The contributions of this article are twofold. The dialectometric results are presented in the previous section. On the methodological side, we demonstrate how multivariate methods can be used to gain insight on dialect variation, and how data can be preprocessed to overcome the problem of uneven sampling.

A central problem in our data is the uneven sampling of the municipalities. This is not an uncommon problem in this type of applications. This, together with the large size of the dataset, means that approaches like clustering and multidimensional scaling do not work when applied directly to the raw data. To deal with this we computed the truncated singular value decomposition of the data matrix, a technique often used in text document analysis. In addition, we removed the contribution of the first singular vector, as this contained the information of the sampling rate and very little else. This approach summarizes the data while reducing the effects of uneven sampling. To this data set we have applied various methods, which now yield meaningful results, unlike in the case of applying them directly to the raw data. An additional benefit is the speed-up of computations, due to the reduced size of the data matrix. Note that the use of this preprocessing scheme is by no means limited to the data analysis methods presented here.

Clustering this data gives a partitioning of the municipalities, which as an approach is similar to the traditional way of defining dialect regions. What is new is that our partitioning relies only on quantitative analysis of lexical variation. In our experience the K-means algorithm with the cosine distance measure seemed better suited for this task than hierarchical clustering.

However, clustering might not be the best way to view dialect variation. Looking at the results of a component analysis might in fact be more informative. After all, borders between different dialects are usually not steep, but vocabulary changes gradually. In analysing this, the various methods take rather different approaches. The principal components capture the main directions of variation, and allow us to view them separately. However, it should be kept in mind that each successive principal component disregards all of the variation already contained in the previous ones, and because of this the components soon become hard to interpret. On the other hand, independent components and those resulting from FA can be more easily interpreted as a diffusion pattern around a dialectal centre, although here too one should be careful.

Note that all methods used here produce geographically meaningful regions. This happens because municipalities which are linguistically close to each other are also geographically close to each other. The methods themselves do not use any geographical information, even if this information would be available. It is also worth pointing out that the results correspond relatively well to earlier views on

Finnish dialects. This indicates that the methods are indeed useful; at the same time, it can also be viewed as an independent confirmation of the eariler views. Nevertheless, there are also some differences, but whether this is because of the different diffusion patterns of lexical and phonological features or a difference between a synchronic and diachronic viewpoint is still somewhat open. However, a thorough linguistic interpretation is beyond the scope of this article.

Finally, it is worth pointing out that some of the interesting results here come from a combination of different methods. For example, we have first used K-means clustering to detect dialect regions, and then used multidimensional scaling on this data to gain insight on the relative distances of these dialects.

Though we have here concentrated on analysing lexical variation, the same methods can be used to analyse phonological and morphological features. It would indeed be interesting to compare results obtained that way to the results presented here.

References

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, ACL, pp. 89–96.

Berry, M. W., Drmač, Z., and Jessup, E. R. (1999). Matrices, vector spaces, and information retrieval. *SIAM Review*, **41**(2): 335–62.

Bloomfield, L. (1933). *Language*. New York: Henry Holt & Co.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Learning*, **1**(2): 224–7.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, **41**(6): 391–407.

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

Embleton, S. and Wheeler, E. S. (2000). Computerized dialect atlas of finnish: dealing with ambiguity. *Journal of Quantitative Linguistics*, **7**(3): 227–31.

Gooskens, C. and Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, **16**: 189–207.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, Massachuetts: MIT Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer Series in Statistics.

Heeringa, W. and Nerbonne, J. (2002). Dialect areas and dialect continua. *Language Variation and Change*, **13**: 375–398.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**: 417–41, 498–520.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, **2**(3): 94–128.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. New York: John Wiley & Sons.

Itkonen, T. (1965). Proto-Finnic Final Consonants, In *Suomalais-Ugrilaisen Seuran toimituksia*, No. 138:1. Suomalaisen Kirjallisuuden Seura.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall, Englewood Cliffs.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New Jersey: John Wiley & Sons.

Kessler, B. (1995). Computational dialectology in irish gaelic. In *Proceedings of the European Associaton for Computational Linguistics*. Dublin, pp. 60–7.

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

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**: 259–84.

Leino, A., Hyvönen, S., and Salmenkivi, M. (2006). Mitä murteita suomessa onkaan? murresanaston levikin kvantitatiivista analyysiä. *Virittäjä*, **110**(1): 26–45.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**(8): 707–10.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J. (eds), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, University of California Press, pp. 281–97.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

Mielikäinen, A. (1991). *Murteiden murros. Levikkikarttoja nykypuhekielen piirteistä*, In *Jyväskylän yliopiston suomen kielen laitoksen julkaisuja*, No. 36. University of Jyväskylä.

Moisl, H. and Jones, V. (2005). Cluster analysis of the newcastle electronic corpus of tyneside english: a comparison of methods. *Literary and Linguistic Computing*, **20**: 125–46.

Nerbonne, J. (2003). Linguistic variation and computation. In *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 3–10.

Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing*, **21**(4): 463–75.

Nerbonne, J. and Heeringa, W. (2001). Computational comparison and classification of dialects. *Dialectologia et Geolinguistica*, **9**: 69–83.

Nerbonne, J., Heeringa, W., and Kleiweg, P. (1999). Edit distance and dialect proximity. In Sankoff, D. and Kruskal, J. (eds), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Stanford, CSLI Press, pp. v–xv.

Palander, M., Opas-Hänninen, L. L., and Tweedie, F. (2003). Neighbours or enemies? competing variants causing differences in transitional dialects. *Computers and the Humanities*, **37**: 359–72.

Paunonen, H. (1991). Till en ny indelning av de finska dialekterna. *Fenno-Ugrica Suecana*, **10**: 75–9.

Porthan, H. G. (1870 [1801]). De praecipuis dialectis linguae fennicae. In Heimbürger L. (ed.), *Henrici Gabrielis Porthan opera selecta*, Vol. IV, Finska Litteratur-Sällskapet, pp. 317–31.

Rapola, M. (1961). Johdatus suomen murteisiin, In *Tietolipas*, No. 4. Suomalaisen Kirjallisuuden Seura, 2nd revised edn.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, **24**(5): 513–23.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Savijärvi, I. and Yli-Luukko, E. (1994). Jämsän äijän murrekirja. In *Suomalaisen Kirjallisuuden Seuran toimituksia*, No. 618. Suomalaisen Kirjallisuuden Seura.

Schone, P. and Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of Conference on Natural Learning*. ACL, Lisbon, pp. 67–72.

Shackleton, R. G. (2005). English-american speech relationships: a quantitative approach. *Journal of English Linguistics*, **33**(2): 99–160.

Strandberg, J. (2004). Ei sanat salahan joua: fennistiikan murteenkeruun historiaa 1868–1925, Master's thesis, University of Helsinki, Department of Finnish.

Theodoridis, S. and Koutroumbas, K. K. (2003). *Pattern Recognition*. San Diego, USA: Academic Press.

Tuomi, T., ed. (1989). Suomen murteiden sanakirja. Johdanto, In *Kotimaisten kielten tutkimuskeskuksen julkaisuja*, No. 36. Kotimaisten kielten tutkimuskeskus.

Vhael, B. G. (1733). *Grammatica Fennica*, Johan Kiämpe. Facsimile: Vanhat kielioppimme. Suomalaisen Kirjallisuuden Seura 1968.

Warelius, A. (1848). Bidrag till Finlands kännedom i ethnographiskt hänseende, *Suomi*, **7**: 47–130.

Wiik, K. (2004). Suomen murteet. Kvantitatiivinen tutkimus, In *Suomalaisen Kirjallisuuden Seuran toimituksia*, No. 987. Suomalaisen Kirjallisuuden Seura.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, **16**(3): 645–78.

Note

1 In practice Ward's method seems to produce meaningful results with other distance measures as well, it just interprets the distances as Euclidean ones. We tried it for the whole data set with the cosine distance measure with much better results than using the Euclidean; Ward's method has also been used together with the Levenshtein distance in (Heeringa and Nerbonne, 2002).