Comparison of the Methods of Self-Organizing Maps and Multidimensional Scaling in Analysis of Estonian Emotion Concepts

Toomas Kirt
Tallinn University of Technology
Tallinn
Estonia
Toomas.Kirt@mail.ee

Ene Vainik
Institute of the Estonian Language
Tallinn
Estonia
ene@eki.ee

Abstract

Self-organizing map (SOM) and multidimensional scaling (MDS) are the methods of data analysis that reduce dimensionality of the input data and visualize the structure of multidimensional data by means of projection. Both methods are widely used in different research areas. In the studies of emotion vocabulary and other psycho-lexical surveys the MDS has been prevalent. In this paper both of the methods are introduced and as an illustration they are applied to a case study of Estonian emotion concepts. There is a need to introduce some new methods to the field because exploiting only one analytical tool may tend to reveal only specific properties of data and thus have an unwanted impact on the results.

1 Introduction

Human's ability to perceive the structure of multidimensional data is limited and some methods are needed to reduce the dimensionality of data and to reveal its structure. Several methods and techniques of data analysis are used to project multidimensional data into a lower two- or three-dimensional space and to visualize the structure of it. In this paper the methods of self-organizing map (SOM) and multidimensional scaling (MDS) are under discussion.

Some of the researchers have compared the methods of SOM and MDS earlier and outlined both their similarities and dissimilarities (e.g., Kaski, 1997; Duda et al., 2001). Kaski has emphasized their general similarity in respect that both methods tend to reduce dimensionality of observed data and reveal its hidden structure. The two methods differ in the strategy applied to the data. The SOM tries to preserve local neighborhood relations and MDS the interpoint distances between samples.

A hypothesis could be formulated that the way the data are handled in an analytical tool might have an impact on the layout of the results. In order to test this hypothesis the data of present case study – a study of the Estonian concepts of emotion – was analyzed by both SOM and MDS. In the following we will demonstrate the layout of data on both cases and discuss their compatibility.

One of the purposes of the comparison of the two methods is to introduce the method of SOM as relatively unexploited in psycho-lexical studies. Although there are some examples of applying SOM to linguistic data (e.g., Honkela, 1997; Lagus et al., 2002) there are no references to other studies of emotion concepts by the self-organizing maps, yet. In the field of psycho-lexical studies MDS has prevailed so far (e.g., the MDS based Geneva Emotion Wheel (Scherer, 2005)), despite SOM's great popularity in several areas of data analysis (Kohonen, 2000).

In the first part of the paper the two methods are introduced. In the second part of the paper the survey of Estonian emotion concepts is used as an example to demonstrate the similarities and differences of the methods.

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 113–120

2 The Self-Organizing Map

The self-organizing map (Kohonen, 1982; 2000) is a tool for the visualization of high-dimensional data. It projects nonlinear relationships between high-dimensional input data into a two-dimensional output grid, named also a map. The self-organizing map is an artificial neural network that uses an unsupervised learning algorithm – it means there is no prior knowledge how input and output are connected.

To describe how the process for creating the self-organizing map works let assume, that we have input data as a set of sample vectors x. It is also called an input space. The output of the self-organizing map is a grid of vectors m_i that have the same number of elements as the sample vector x. Initially all the vectors of the output grid are initialized randomly.

The algorithm of SOM has two main basic steps that are repeated a number of times. First a random sample vector x(t) is chosen and compared with all the output vectors m_i to find closest unit c on the output grid that has a minimum distance $d(x - m_i)$ with a sample vector x. Secondly this best matching or winning vector and its neighborhood are changed closer to the sample vector. The formula for learning process is as follows:

$$m_i(t+1) = m_i(t) + a(t)\, h_{ci}(t)(x(t) - m_i(t)).$$

Where a(t) is learning rate factor and $h_{ci}(t)$ – neighborhood function at the time step t. During the learning process the learning rate and the neighborhood function are shrinking. The learning process results in an ordered output where similar sample vectors are projected as closely located units on the map.

For visualization of the self-organizing map an Unified distance matrix (U-matrix) is used (Ultsch, 1993). The U-matrix presents the distances between each map unit by color coding. The light color corresponds to a small distance between two map units and the dark color presents a bigger difference between the map units. The points on the output map that are on the light area belong to the same group or cluster and the dark area shows the borders between the clusters.

To illustrate the behavior of the SOM the matrix of distances between Estonian cities is used. The input data consists of distances between 59 Estonian cities. The initial distance matrix is downloaded from the web page of the Estonian Road Administration[1]. From the distance matrix the relative coordinates are calculated. The coordinate matrix is two-dimensional and therefore it is useful to see, how a method transforms the original data. The analysis is performed by the SOM toolbox ver 2.0 for Matlab[2].

The output of the SOM is presented in Figure 1. The map retains Estonian original topological structure in general terms, despite the fact that the eastern side of Estonia is projected on the top of the map. The cities that are close to each other in the real map are projected on the close map units. The color coding also gives some insight into distances between the cities and it is possible to identify regions where the density of population is higher. The local neighborhood is retained, but it is difficult to fully identify the map with the real map of Estonia.

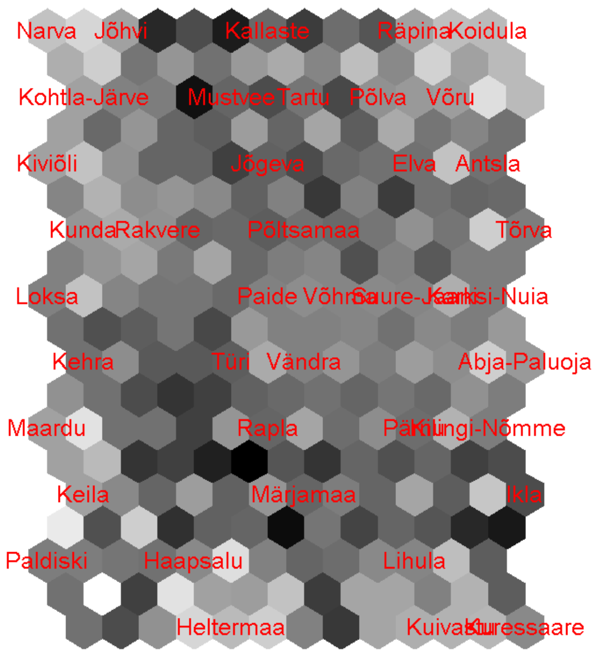

Figure 1. The SOM of Estonian Cities.

3 Multidimensional Scaling

The method of multidimensional scaling (MDS) is a set of related statistical techniques often used in data visualization for exploring proximities in data. The goal of the method is to project data points as points in some lower-dimensional space so that the

[1] Downloaded from http://www.mnt.ee/
[2] Downloaded from http://www.cis.hut.fi/projects/somtoolbox/

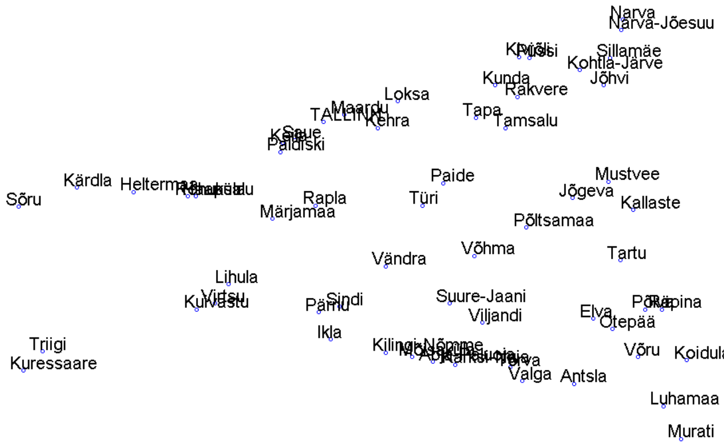

Figure 2. The MDS of Estonian Cities.

distances between the points correspond to the dissimilarities between the points in the original space as closely as possible. Such representation is valuable for gaining insight into the structure of data. MDS can be used as a method of reducing the dimensionality of the data and revealing the dissimilarity between the samples.

MDS is said to be metrical if it based on measured proximities and nonmetrical when the proximities are based on judgment (Jobson, 1992). The original method of MDS was metric (Torgerson, 1958). In current paper the analysis is based on nonmetrical data and therefore the nonmetric MDS is used. The data is analyzed by the statistical software package SPSS and the ALSCAL algorithm created by Takane et al. (1977).

There are n sample vectors $x_1,...,x_n$ and the distance between original samples i and j is g_{ij}. The y_i is the lower-dimensional representation of x_i and the distance between projected samples i and j is d_{ij}. The aim of the MDS method is to find a configuration of image points $y_1,...,y_n$ in a lower dimensional space for which the distances d_{ij} between the samples are as close as possible to the corresponding original distances g_{ij} so that the dissimilarities between the samples are retained as well as possible. Because it is impossible to find a configuration for which $d_{ij} = g_{ij}$ for all i and j, certain criteria are needed whether the result is good enough.

The interdistance matrix of Estonian cities is used again to illustrate the method of MDS (Figure 2). As it can be noticed the result resembles Estonian map despite the fact that some cities in the Northwest and Southwest are projected closer than they are in the real map. It can be caused by the well known “horseshoe effect” that is common to the multidimensional scaling (Buja and Swayne, 2002).

As we can see from the initial example (Figures 1 and 2) the two methods have their preferences. The SOM is good, if the data is represented as coordinates and local relations between the samples are important. The MDS is oriented to reveal the structure of metric distances between the samples and it reveals the overall picture of the data.

4 Study of Estonian Emotion Concepts

The purpose of the case study was to discover the hidden structure of the Estonian emotion concepts and whether it depended on how the information about concepts was gathered. According to the theory of conceptual spaces (Gärdenfors, 2000), the level of conceptual representations of emotions is assumed to be intermediate in abstractness between the levels of purely linguistic (symbolic) and subconceptual representation which is related to emotional experience. In the experiment these two levels of emotion knowledge (lexical and experiential) were used to approach the intermediate level of concepts. Two lexical tasks were designed that provided information about emotion concepts either through their relation to the episodes of emotional experience or through semantic interrelations of emotion terms (synonymy and antonymy).

4.1 Subjects and Procedures

The inquiry was carried out in written form, in 2003, in Estonia. The number of respondents was 100 (50 men and 50 women), aged from 14 to 76 (*M* = 40.2, *SD* = 18.61), all native speakers of Estonian. The selection of concepts to be included in the study (N=24) was based on the results of tests of free listings (Vainik, 2002), word frequencies in the corpora, and a comparison with word lists used by some earlier studies of Estonian emotion terms. We believe that the selected lexical items form a small but representative set of the core of the emotion category of Estonian lexicon, sufficient for comparing the structures of emotion concepts, which emerge from the two different lexical tasks.

In the first task the participants had to evaluate the meaning of every single word against a set of seven bipolar scales, inspired by Osgood's method of semantic differentials (Osgood et al., 1975). The "semantic features" measured with polar scales drew qualitative (unpleasant vs. pleasant), quantitative (strong vs. weak emotion, long vs. short in duration), situational (increases vs. decreases action readiness, follows vs. precedes an event), and interpretative distinctions (felt in the mind vs. body, depends mostly on oneself vs. others). The original bipolar scales were transformed from having +/- values into positive scales of 7–1, starting from 7 as the maximum value of the dominant or default feature, over 4 pointing to the irrelevance of the scale, and up to 1 as the minimum value (corresponding to the maximum of the opposite feature).

The second task was a free listing task (Corbett and Davies, 1997). Participants were provided with a blank space to write down as many synonyms and antonyms as came to mind for every presented item. The task eliciting similar concepts resulted in 4068 lexical items and the task eliciting opposite concepts resulted in 3694 lexical items. Before the analysis with SOM and MDS the information was first quantified. The words listed as similar or opposite were characterized by their indices of relative cognitive salience (Sutrop, 2001). The index which takes into account both frequency and mean position of a term was calculated for every word mentioned by at least three persons. Out of total 488 relations only 219 with indices greater than or equal to the average (S_{ave} = .07), were subsequently processed with SOM and MDS

4.2 Results of Task 1 and Task 2

In the first task the data pool of all answers to the 24 concepts on the 7 joint scales was processed. So a vector consisting of 700 answers represented each word. In the second task the words were described by a vector in length of 219 representing values of the index of relative cognitive salience.

Figure 3 and 4 present the structure of Estonian emotion concepts according to the results of the first task. The translations and locations of words on the SOM are given in the following Table 1. The MDS was created with translations only.

The SOM of the first task appears as a bilaterally symmetrical representation. The positive emotion concepts tend to gather to the upper part of the graph and the words referring to negative emotions to the lower part of the graph. Thus, the main organizing dimension of the representation, which extends the shape of the SOM map in one direction, appears to be negativeness and positiveness of the concepts. There is a darker area in the middle, which clearly separates these two clusters. One concept, *ärevus* 'anxiety', is located outside of these two clusters. Apparently it is identifiable neither as positive nor negative or having conflicting specifications in respect of affiliation. As the anticipatory states (*hirm* 'fear', *erutus* 'excitement', *mure* 'concern') are gathered to the right edge of the graph, the scale follows vs. precedes an event seems to function as an additional less important dimension. There is, however, no darker area on the SOM separating the extremes of this dimension.

The MDS represents concepts on the circle. By shape it resembles the circumplex model proposed by Russell (Russell, 1980; Russell et al., 1989). The MDS presents also a clear distinction between the positive and negative concepts on the horizontal scale – the more negative the concepts the more left they are situated and the positive concepts are situated on the right-hand side, accordingly. In MDS, too, the concept of *ärevus* 'anxiety' occurs as ambivalent between positive and negative concepts, and so does *kaastunne* 'pity, compassion'.

Figure 3. The SOM of the First Task.

Table 1. Location of Words on the SOM of the First Task.

enthusiasm happiness joy	pleasure fun love	passion
		excitement desire
surprise pride		
		anxiety
pity		
rage envy		concern
anger		
guilt disappointment contempt	sadness shame oppression	fear

There is another dimension that distinguishes the concepts on vertical scale: the states perceived as event preceding are situated on the upper part of the circle and the states perceived as following some event are situated in the bottom. According to the MDS presentation the concept *masendus* ‘oppression’ can be regarded as not clearly preceding nor following its eliciting event.

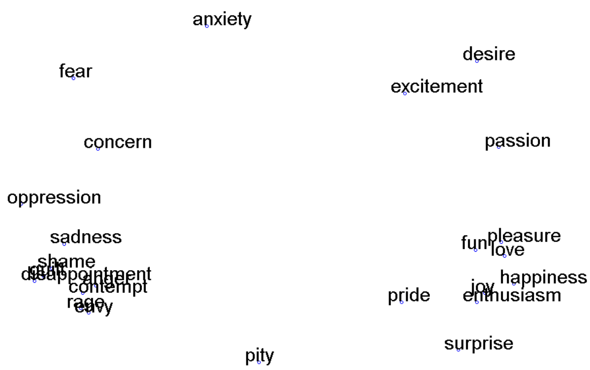

Figure 4. The MDS of the First Task.

The results of the first task characterize how the conceptual organization of emotion emerged from subconceptual and experiential level of knowledge in Gärdenfors’s model (2000). It can be seen, that the two methods resulted in very similar layouts, except the orientation of the dimensions and the way of discriminating the groups.

Figures 5 and 6 (and Table 2) present the structure of the Estonian emotion concepts according to the results of the second task of the survey. This task addressed the most abstract and symbolic level of representation of emotion knowledge, according to the Theory of conceptual spaces (Gärdenfors, 2000), which was accessed through the semantic interrelations of emotion terms in our task.

On the SOM of the second task also a general vertical alignment of positive (bottom) versus negative (top) concepts is observable. There is a remarkably darker row of nodes aligned horizontally, separating those two categories of unequal size. The concepts have self-distributed into three clusters, though, as in the upper part of the graph there is a diagonally located darker area excluding the cluster of concepts in the uppermost right corner. One node containing two concepts *iha* ‘desire’ and *kirg* ‘passion’ are standing outside the clusters not belonging to any of them.

This SOM does not coincide with the SOM of the Task 1. Instead of two we have three clearly distinguishable clusters here. This lets us to conclude that the organization of emotion concepts is slightly different while emerging from the data about the relations of similarity and oppositeness. The SOM layouts thus occur to support the hypothesis of the case study about the plausibility of differences in conceptual organization due to the way the data about concepts is gathered.

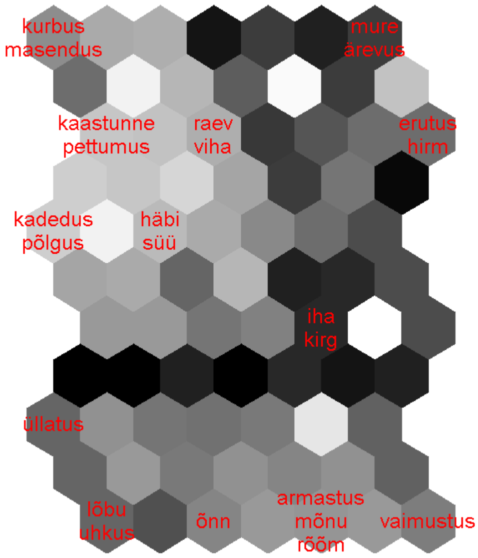

Figure 5. The SOM of the Second Task.

Table 2. Location of Words on the SOM of the Second Task.

sadness oppression			concern anxiety
pity disappointment	rage anger		excitement fear
envy contempt	shame guilt		
		desire passion	
surprise			
fun pride	happiness	love pleasure joy	enthusiasm

The MDS of the second task, on the other hand, retained the circular structure and there might be seen the horizontal alignment of positive (right-hand side) versus negative (left-hand side) concepts on the graph, as well as the vertical alignment of event preceding states (the upper part) versus the event following states (the lower part of the graph).

At the first glance the result of Task 2 as analyzed by MDS is very similar to the result of Task 1 except that the locations of *kaastunne* ‘pity’ and *vaimustus* ‘enthusiasm’ do not fit. This result leads us to two possible conclusions. First, we can conclude that the way the information about emotion concepts was gathered had no or only nonsignificant impact on their emergent structure, which proves the invalidity of our hypothesis of the case study. On the other hand, we can conclude that the method of MDS tends to generalize the results to fit a circular solution best presented by two crossing dimensions.

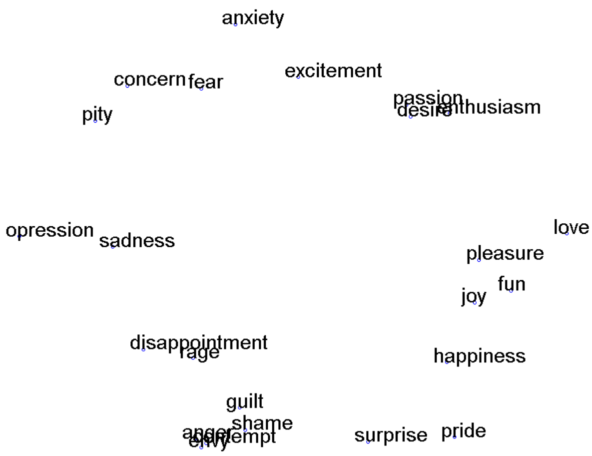

Figure 6. The MDS of the Second Task.

However, even on the circular arrangement there are actually three groups of concepts visible, especially with the prior knowledge from the SOM analysis. On the bottom right there is a cluster of positive concepts, the cluster of negative ones is situated on the bottom left and on the top there are concepts that might be described mostly by their quality as event preceding states. These three clusters are partly compatible with these three described on the SOM of the Task 2 (Figure 5).

5 Discussion

In previous section two tasks of differently accessed semantics of the Estonian emotion terms were compared and two methods of data analysis were applied. As a result, both methods gave us a general understanding what are the main dimensions that distinguish emotion concepts and revealed that there is clear distinction between positive and negative concepts. In the first task both methods distinguished two groups of concepts and in the second task one additional cluster emerged. The level of abstractness at which emotion knowledge was accessed in the tasks (subconceptual and experience-related vs. symbolic and lexicon-related) turned out as critical while SOM was used and nonsignificant while MDS was used. The hypothesis of the case study was thus proven only in the case of using SOM. With this conflicting result, however, is proven the main hypothesis of our present study. Namely, the way the data was handled in an analytical tool turned out to have an impact on the layout of the results.

Comparing the results of analysis of linguistic data SOM formed clearly separable clusters and MDS projected data on the circle. Supposedly, MDS presented the overall distances between the samples and therefore the extremity of dominant positive negative scale became dominant in both cases and the overall layout of the results occurred as the same - circular. At the same time the SOM gives an overview of local relations between concepts and forms local clusters. However, even the projection of local relationships between the samples gave us the insight that there is the division between the positive and negative concepts.

In the case the data was gathered from the task relying on the procedure of the Osgood’s semantic differential or alike, the two methods revealed very similar results. In the case the data was gathered by assessing concept similarity and oppositeness the layouts of MDS and SOM seem somehow differently. It is probably the point where the different strategies used in the analytical tools turn out as critical. MDS uses a strategy to keep most dissimilar samples as apart as possible (it preserves the distances) and SOM uses the strategy to keep the most similar samples together (it preserves the neighborhood relations). The data of the Task 2 contained data about both assessed concept similarity (a tendency to interpret similar concepts as situated close to each other) and about oppositeness (a tendency to interpret most dissimilar samples as most apart in a hypothetical conceptual space (Gärdenfors, 2000)). Thus the construal of the Task 2 might have made it sensitive to the procedures used in the analytical tool.

While analyzing linguistic data containing information about concept similarities and dissimilarities it might be useful not to be grounded in just one analytical tool, because MDS gave similar circular structure as a result of both tasks. When some additional knowledge was acquired from the SOM analysis, a more complicated structure within the data was revealed. The interpretation of the results may depend on the interpreter – his or her thoroughness and in more general what he or she wants or supposes to see.

6 Conclusions

In the present paper the results of analysis of Estonian emotion concepts by two methods — the self-organizing maps and multidimensional scaling — were compared. Both methods gave us a general understanding what are the main dimensions distinguishing emotional concepts and revealed a clear distinction between positive and negative ones. Both methods also demonstrated their peculiarities due to the different strategies used in their procedures of data handling. Although both methods reveal the dominant dimensions describing the data, SOM stresses more on the local similarities and distinguishes clearly groups within the data. MDS reveals global dissimilarities between the samples and some background information is needed to distinguish groups. Our conclusion would be that exploiting only one analytical tool may tend to reveal only specific properties of data and thus have an unwanted impact on the results.

Acknowledgement

The study was supported by ETF grant 7149.

References

Buja A. and Swayne D. F. 2002. Visualization Methodology for Multidimensional Scaling. *Journal of Classification*, 19: 7-43.

Corbett G. G. and Davies I. R. L. 1997. Establishing basic color terms: Measures and techniques. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 197–223). Cambridge University Press, Cambridge.

Duda R. O., Hart P. E. and Stork D. G. 2001. *Pattern classification* (2nd ed.). John Wiley & Sons, New York.

Gärdenfors P. 2000. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, London.

Honkela T. 1997. Learning to understand—General aspects of using self-organizing maps in natural language processing. In D. Dubois (Ed.), *Computing anticipatory systems* (pp. 563–576). American Institute of Physics, Woodbury, NY.

Jobson J. D. 1992. *Applied multivariate Data Analysis*, Vol. II. Springer, New York.

Kaski S. 1997. Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82*. Helsinki University of Technology, Finland.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43: 59–69.

Kohonen T. 2000. *Self-organising maps* (3rd edition), Springer, Berlin.

Lagus K., Airola A. and Creutz M. 2002. Data analysis of conceptual similarities of Finnish verbs. In W. D. Gray and C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 566–571). Lawrence Erlbaum, Hillsdale, NJ.

Osgood C. E., Suci G. J. and Tannenbaum, P. H. 1975. *The Measurement of Meaning*. University of Illinois Press, Urbana and Chicago.

Russell J. A. 1980. A circumflex model of affect. *Journal of Personality and Social Psychology*, 39: 1161–1178.

Russell J. A., Lewicka M. and Niit T. 1989. A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57: 848–856.

Scherer K. R. 2005. What are emotions? And how should they be measured? *Social Science Information*, 44(4): 695–729.

Sutrop U. 2001. List task and a cognitive salience index. *Field Methods*, 13: 289–302.

Takane, Y., Young, F. W., and de Leeuw J. 1977. Nonmetric individual differences multidimensional scaling: an alternating least square method with optimal scaling features. *Psychometrika*, 42: 7-67.

Torgerson, W. S. 1958. *Theory and methods of scaling*. Chapman & Hall, London.

Ultsch, A. 1993. Knowledge extraction from self-organizing neural networks. In O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and Classification* (pp. 301–306). Springer, Berlin.

Vainik, E. 2002. Emotions, emotion terms and emotion concepts in an Estonian folk model. *Trames*, 6(4): 322–341.