Subjects on Objects in Contexts: Using GICA Method to Quantify Epistemological Subjectivity

Timo Honkela, Juha Raitio,
Krista Lagus and Ilari T. Nieminen
Aalto University School of Science
Dep't of Information and Computer Science
P.O.Box 15400, FI-00076 AALTO, Finland
Email: first.last@aalto.fi

Nina Honkela
University of Helsinki
P.O.Box 18
FI-00014 Helsinki
Email: nina.honkela@helsinki.fi

Mika Pantzar
National Consumer Research Center
P.O.Box 5
FI-00531 Helsinki
Email: mika.pantzar@ncrc.fi

***Abstract*—A substantial amount of subjectivity is involved in how people use language and conceptualize the world. Computational methods and formal representations of knowledge usually neglect this kind of individual variation. We have developed a novel method, Grounded Intersubjective Concept Analysis (GICA), for the analysis and visualization of individual differences in language use and conceptualization. The GICA method first employs a conceptual survey or a text mining step to elicit from varied groups of individuals the particular ways in which terms and associated concepts are used among the individuals. The subsequent analysis and visualization reveals potential underlying groupings of subjects, objects and contexts. One way of viewing the GICA method is to compare it with the traditional word space models. In the word space models, such as latent semantic analysis (LSA), statistical analysis of word-context matrices reveals latent information. A common approach is to analyze term-document matrices in the analysis. The GICA method extends the basic idea of the traditional term-document matrix analysis to include a third dimension of different individuals. This leads to a formation of a third-order tensor of size** $\text{subjects} \times \text{objects} \times \text{contexts}$**. Through flattening into a matrix, these subject-object-context (SOC) tensors can again be analyzed using various computational methods including principal component analysis (PCA), singular value decomposition (SVD), independent component analysis (ICA) or any existing or future method suitable for analyzing high-dimensional data sets. In order to demonstrate the use of the GICA method, we present the results of two case studies. In the first case, GICA of health-related concepts is conducted. In the second one, the State of the Union addresses by US presidents are analyzed. In these case studies, we apply multidimensional scaling (MDS), the self-organizing map (SOM) and Neighborhood Retrieval Visualizer (NeRV) as specific data analysis methods within the overall GICA method. The GICA method can be used, for instance, to support education of heterogeneous audiences, public planning processes and participatory design, conflict resolution, environmental problem solving, interprofessional and interdisciplinary communication, product development processes, mergers of organizations, and building enhanced knowledge representations in semantic web.**

I. Introduction

Often we take for granted that we are able to understand each other. It is the occasional clear failure that allows us to see that understanding language is often difficult.

In making the connection between a word and its typical and appropriate use, we humans rely on a long learning process. The process is made possible and guided by our genetic make-up, but its success essentially requires extensive immersion to a culture and contexts of using words and expressions. To the extent that these contexts are shared among individual language speakers, we are then able to understand each other. When our learning contexts differ, however, differences in understanding the concepts themselves arise and subsequent communication failures begin to take place.

It is obvious that if the context of learning has been completely different, i.e., if two persons have learned different languages, the basis for mutual understanding through exchanging linguistic expressions is very limited or even non-existent. Self-evidently, without an access to gestures or an external context it is not possible to know what "Ble mae'r swyddfa bost agosaf?" or "Non hurbilen dagoen postetxean da?" means unless one has learned Welsh or Basque language. This example can naturally be extended to less trivial cases as well. Considering the readers of this article it is fair to assume that every one of them speaks English. Nevertheless, it is difficult for most to understand expressions like "a metaphyseal loading implant employes a modified mechanoregulatory algorithm" or "bosonic fields commute and fermionic fields anticommute" unless one is an expert in a particular area of medicine or physics. Even expressions in everyday informal language such as "imma imba, lol" can seem obscure if one is not familiar with the youth language in the internet. In addition to these kinds of clear-cut cases, there are more subtle situations in which two or several persons think that they understand each other even though they actually do not. It seems realistic to think that a person assumes that others understand her when she says "this is not fair", "do you like me?", "I saw a small house", or "that country is democratic". However, it is far from guaranteed that the others would actually interpret words "fair", "like", "small" or "democratic" in the same way as the speaker.

In this paper, building on previous work [1], [2], [3], [4], we present a methodological innovation that aims to improve a) mutual understanding in communication, and b) inclusion of stakeholder concerns in complex decision-making contexts. The proposed method builds on 1) an understanding of the grounded nature of all concepts, and the dynamic and subjective nature of concept formation and use; and 2) the recognition that the best way to elicit and represent such concepts is one that combines elements from qualitative case research and quantitative learning methods. We call this method *Grounded Intersubjective Concept Analysis (GICA)*. The word 'grounded' refers to both the qualitative method of Grounded Theory [5] and to the idea of the embodied grounding of concepts in human experience [6]. Areas of application for the GICA method are, for instance, public planning processes, environmental problem solving, interdisciplinary research projects, product development processes, and mergers of organizations.

A. Contextuality and subjectivity

It is commonplace in linguistics to define semantics as dealing with prototypical meanings, whereas pragmatics would be associated with meanings in context. For our purposes, this distinction is not relevant, however, since interpretation of natural language expressions always takes place in some context, usually even within multiple levels of context including both linguistic and extra-linguistic ones. In the contrary case, i.e., when an ambiguous word such as "break" appears alone without any specific context one can only try to guess which of its multiple meanings could be in question. If there is even a short contextual cue — "break the law", or "have a break", or "how to break in billiards" – it is usually possible to arrive at a more accurate interpretation. Also the extralinguistic context of an expression usually helps in disambiguation.

In some cases, the interpretation of expression can be numerically quantified and thus more easily compared. For instance, the expression "a tall person" can be interpreted as a kind of measure of the height of the person. The interpretation of 'tallness' can be experimentally studied in two ways. Either one can be asked to tell the prototypical height of a person that is tall, or one can tell whether different persons of some height are tall or not (maybe associated with some quantifiers such as "quite" or "very"). Sometimes this kind of quantification is conducted using the framework of fuzzy set theory [7]. However, consideration of the tallness of a person is only the tip of an iceberg of the complexity of interpretation. A small giraffe or building is usually higher than a tall person. A person who is 5 feet or 1 meter 52 centimeters is not prototypically considered tall – unless a young child is in question. Also many other contextual factors influence the interpretation such as gender, historical time (people used to be shorter hundreds of years ago), and even profession (e.g., basketball players versus fighter pilots).

The tallness example also provides a view on subjectivity. If we ask from a thousand people the question "How tall is a tall person?", we receive many different answers, and if we ask "If a person is x cm tall, would you call him/her a tall person?", the answer varies among respondents. The distribution of answers to such questions reflects the individual variation in the interpretation of 'tall'. If the pattern in question is more complex and a number of context features are taken into account, the issue of subjective models becomes even more apparent, unless it is assumed that such information for interpretation (linking language with perceptions) would be genetically determined which clearly appears not to be the case.

Another simple example on subjectivity is found in naming colors. Differences and similarities in color naming and color concepts in different languages have been studied carefully (see e.g. [8], [9]). In addition, unless prototypical colors such as pure black, white, red, green, etc. are chosen, individual people tend to name a sample color in different ways. What is dark blue for someone, may be black to someone else, and so on. A similar straightforward illustration of subjectivity of interpretation is the naming of patterns.

It is important to note that the kind of subjectivity discussed above is usually not dealt with in computational or formal theories of language and understanding. On the other hand, this phenomenon is self-evident for practitioners in many areas of activity as well as in relation to practice oriented fields in the humanities. However, subjectivity has been difficult to quantify. In this paper, we introduce a method that is meant to make the subjectivity of interpretation and understanding explicit and visible even in non-trivial cases. The topic of this paper is highly interdisciplinary. Methodologically, we build a framework in which we mainly rely on existing tools used in statistical machine learning and neural networks research. The underlying theoretical issues are related to linguistics, cognitive science, psychology and philosophy of language, and the main application areas are related to social sciences, including organizational research, as well as information system design.

B. Analyzing semantic subjectivity

For the most part, people do not seem to be aware of the subjectivity of their perceptions, concepts, or world views. Furthermore, one might claim that we are more typically conscious of differences in opinions, whereas differences in perception or at conceptual level are less well understood. It is even possible that to be able to function efficiently it is best to mostly assume that my tools of communication are shared by people around me. However, there are situations where this assumption breaks down to a degree that merits further attention. An example is the case when speakers of the same language from several disciplines, interest groups, or several otherwise closely knit cultural contexts come together to deliberate on some shared issues.

The background assumption of the GICA method is the recognition that although different people may use the same word for some phenomenon, this does not necessarily mean that the conceptualization underlying this word usage is the same; in fact, the sameness at the level of names may hide significant differences at the level of concepts. Furthermore,

there may be differences at many levels: experiences, values, understanding of the causal relationships, opinions and regarding the meanings of words. The differences in meanings of words are the most deceptive, because to discuss any of the other differences, a shared vocabulary which is understood in roughly the same way, is necessary. Often a difference in the meanings of used words remains unrecognized for a long time; it may, for instance, be misconstrued as a difference in opinions. Alternatively, a difference in opinions, or regarding a decision that the group makes, may be masked and remain unrecognized, because the same words are used seemingly in accord, but in fact in different meanings by different people. When these differences are not recognized during communication, it often leads to discord and unhappiness about the end result. As a result, the joint process may be considered to have failed in its objectives.

It is worth noting that our work differs considerably from the research in which subjectivity in opinion and its representation in language is in focus (see e.g. [10], [11]).

Mustajoki [12] presents a model of miscommunication based on careful linguistic observations. The underlying insights and motivation of his work resemble to a large extend this article as well as the model presented in [3]. Mustajoki concludes that even in the scientific literature on failures in communication different terms are occasionally used to describe similar matters and researchers also tend to use the same terms with different meanings. In this article, we do not aim to review the research on miscommunication but refer to [12], [13] as good overviews. In the following, we present as our contribution a division into two main types of problems.

Undiscovered meaning differences can cause two types of problems. The first type is *false agreement*, where on the surface it looks as if we agree, but in fact our conceptual difference hides the underlying difference in opinions or world views. For example, we might all agree that "university A should be innovative" or that "university B should aim at excellence in research and education" but could disagree about what "innovative" or "excellence" means. As another example, we might agree that "we need a taxing system that is fair and encourages people to work" but might be in considerable disagreement regarding the practical interpretation of "fair", "encourages" or even "work".

The second type of problem caused by undiscovered meaning differences is *false disagreement*. If we are raised (linguistically speaking) in different sub-cultures, we might come to share ideas and views, but might have learned to use different expressions to describe them. This may lead to considerable amount of unnecessary argument and tension, in short, surface disagreement, that hides the underlying agreement.

Since a lot of human endeavor when meeting with others seems to deal with uncovering conceptual differences in one way or another, it would be beneficial to have tools which can aid us in the discovery process—tools which might make visible the deeper conceptual level behind our surface level of words and expressions.

II. Methodology

Our aim with the Grounded Intersubjective Concept Analysis method is to devise a way in which differences in conceptualization such as those described above can be made visible and integrated into complex communication and decision making processes.

A. Subjectivity tensors

In the GICA method, the idea of considering some items or objects such as words in their contexts is taken a step further. As we have aimed to carefully show in the introductory section of this paper, subjectivity is an inherent aspect of interpretation. In order to capture the epistemological subjectivity, we add a third dimension to the analysis. Namely, we extend the set of observations, $\mathrm{objects} \times \mathrm{contexts}$, into $\mathrm{subjects} \times \mathrm{objects} \times \mathrm{contexts}$, i.e. we additionally consider what is the contribution of each subject in the context analysis.

The resulting data structure could be called a cuboid. However, we use the terminology of *tensor analysis* and adopt the notation used by Kolda and Bader [14]. The *order of a tensor* is the number of the array dimensions, also known as ways or *modes*. As a GICA dataset is observed under varied conditions of the three modes, these form the ways of the order-three tensor

$$\mathcal{X} \in \mathcal{R}^{S \times O \times C},$$

where S, O, C are the number of values (levels) in ranges $\{s_1, s_2, \ldots, s_S\}$, $\{o_1, o_2, \ldots, o_O\}$ and $\{c_1, c_2, \ldots, c_C\}$ of the categorical variables subject $\mathbf{s}$, object $\mathbf{o}$ and context $\mathbf{c}$, respectively. An element of the tensor, $x_{ijk} \in \mathcal{R}$, is the individual observation under certain levels (s_i, o_j, c_k). $\mathcal{R}$ is the range of the observed variable. The idea of expansion into a three-way tensor is illustrated in Fig. 1.

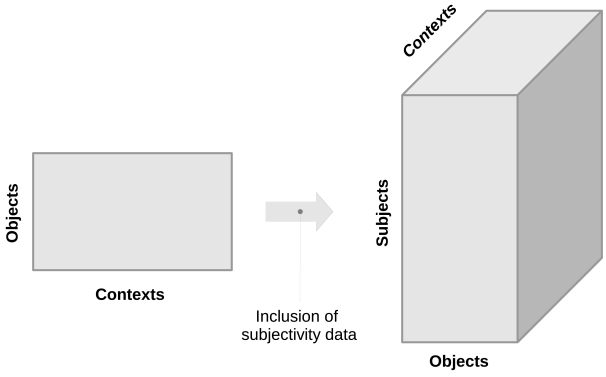

Fig. 1. An illustration of an object-context matrix expanded into a tensor that accounts additionally for subjectivity. In other words, we perform an extension of a $O \times C$-element matrix into a third-order tensor of $S \times O \times C$ elements, where the data concerning different subjects on objects in contexts are included.

As many analysis methods and computations handle matrices, it may be practical to *flatten* (or *unfold*) the GICA data tensor into a GICA data matrix. For this purpose, following [14], we define a *fiber* of a tensor as the subarray consisting

of elements having every index but one fixed. Thus, a way-1 fiber of the data tensor $\mathcal{X}$ is the vector $\mathbf{x}_{:jk}$, where a colon is used to indicate all elements of way 1 and $j \in \{1, 2, \ldots, O\}$, $k \in \{1, 2, \ldots, C\}$. The *way-n matricization* of the data tensor $\mathcal{X}$ can now be defined as rearranging the elements so that the fibers of way n, collected through stepping the two bound indices in breadth-first order, form the columns of the resulting data matrix $\mathbf{X}$. As an example, the way-3 matricization of a GICA data tensor is the matrix of observations with contexts c_k as rows and columns running through all objects o_j for each subject s_i. This matricization is illustrated in Fig. 2. It is important to note that similar operation can be conducted as a way-1 or way-2 matricization. Each of these unfoldings of the tensor gives a specific point of view into the data. In concrete terms, matricizations provide the opportunity to analyze the combinations of subjects and objects, subjects and contexts as well as objects and contexts as columns of matrices, thus avoiding the use of tensor analysis.

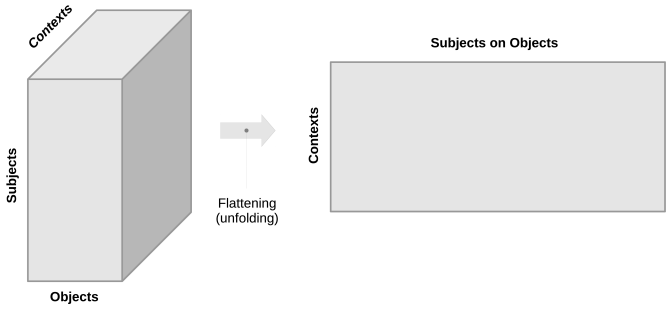

Fig. 2. The $S \times O \times C$-element subjectivity data flattened into a matrix in which each row corresponds to a context and each column to a unique combination of a subject and an object. The number of columns in this matrix is $S \times O$ and the number of rows is C.

B. Obtaining subjectivity data

A central question in GICA is how to obtain the data on subjectivity for expanding an object-context matrix into the tensor that accounts additionally for subjectivity. The basic idea is that for each element in the object-context matrix one needs several subjective evaluations. Specifically, *the GICA data collection* measures for each subject s_i the relevance x_{ijk} of an object o_j in a context c_k, or, more generally, the association x_{ijk} between object and context.

Selection of objects depends on the domain and task in which the GICA method is being used. A typical choice is to concentrate on such objects that are expected to be understood in varying ways among the people or groups of people who are involved. For instance, in the case of a multidisciplinary university conducting strategic planning, the objects could include items such as "science", "productivity", "quality", "design", and "impact". Another selection criterion is the importance of the objects for the task at hand. It is useful to choose objects for which potential underlying disagreement about the meaning and connotations would endanger some central communication or decision making processes.

Contexts serve as representative features of the application area, against which differences in conceptualization between subjects are measured. Thus, *the selection of contexts* should be such that their relation to the objects of interest highlight differences in conceptualization of the objects rather than differences in opinions about the objects. This may be best achieved by selecting contexts that do not directly relate to the possibly conflicting interests of the subjects, and by making contexts as unambiguous as possible.

Empirical methods of psychology and brain research could be considered for obtaining GICA dataset. In brain research, the GICA method could be used to model individual differences in brain-related signals. However, in this article, we present two approaches that are easily usable for anyone who wishes to conduct GICA: 1) conducting a conceptual survey, or 2) applying text mining for a suitable document collection. These approaches are presented in section III including a description of a real world use case for both of them.

C. Interpreting subjectivity data

The target in the interpretation of the subjectivity dataset is to uncover objects, where conceptualization is not commonly shared by the group under study as well as those subjects who are not agreeing on conceptualizations. The *level of agreement* can be measured in terms of the size of the subgroup that share the same conceptualizations. Agreement is small for an individual who's data $x_{i::}$ disagrees on many of the objects with the rest of the group. Further, the level of agreement within a subgroup of subjects can be characterized by the sizes of the subgroups of objects the group members agree on. *The level of subjectivity* can be measured for objects by the number of subgroups of the subjects that share the same conceptualization on the object.

In general, groupings can be revealed by means of cluster analysis, and can be used in inferring about facts useful in the task of interest. In the next sections we aim to show through concrete case examples how cluster analysis can be applied in GICA.

III. Experiments

In the following, we present two experiments that illustrate variants of the GICA method.

A. Conceptual survey of subjectivity

The conceptual survey for obtaining subjectivity data is illustrated with a case study related to wellbeing concepts. The topic was handled in the EIT ICT Labs (http://eit.ictlabs.eu/) activity "Wellbeing Innovation Camp" that took place between 26th and 29th of October 2010 in Vierumäki, Finland. The seminar participants were mainly from Aalto University School of Science and Technology, Macadamia Master's Programme in Machine Learning and Data Mining and from Aalto University School of Art and Design, Department of Design.

In our example, the objects are chosen from the domain of wellbeing. Originally the list consisted of eight objects of interest (wellbeing, fitness, tiredness, good food, stress,

Item	Frequency	Item	Frequency
friends	33	safety	7
health	23	exercise	7
family	23	delicious	7
sleep	14	success	6
music	13	sleeping	6
work	11	relaxation	6
time	10	pressure	6
happiness	10	nutrition	6
depression	10	nature	6
stress	9	home	6
sports	9	wine	5
healthy	9	satisfaction	5
fresh	9	physical health	5
food	9	love	5
darkness	9	hurry	5
sport	8	healthy food	5
freedom	8	deadline	5
traveling	7	bed	5
social interaction	7	...	...

TABLE I
MOST COMMON ITEMS ASSOCIATED BY THE PARTICIPANTS WITH EIGHT TERMS RELATED TO WELLBEING.

relaxation, loneliness and happiness), but at a later stage of the process the list was narrowed down to four objects (relaxation, happiness, fitness, wellbeing).

The next step in the method is to collect a number of relevant contexts towards which the previously collected objects can be reflected. In principle, the context items can be short textual descriptions, longer stories, or even multimodal items such as physical objects, images or videos. The underlying idea is that between the objects and the contexts there is some kind of potential link of varying degree. It is important to choose the contexts in such a manner that they are as clear and unambiguous as possible. The differences in the interpretations of the objects is best revealed if the "reflection surface" of the contexts is as shared as possible among the participants. Therefore, the contexts can include richer descriptions and even multimodal grounding.

The number of objects and contexts determines the overall number of inputs to be given. Naturally, if the number of objects and/or contexts is very high, the task becomes overwhelming to the participants. Therefore the number of objects should be kept reasonable, for instance between 10 and 15, and the number of contexts should be such that the dimensions are enough to bring to the light the differences between the conceptual views of the persons.

In the wellbeing workshop, the participants were asked to list concepts related to eight areas of wellbeing (wellbeing, fitness, tiredness, good food, stress, relaxation, loneliness and happiness). The participants listed 744 terms among which 182 were mentioned by more than one person. Unique items included "homesickness", "handicrafts", "grandma's pancakes", etc. The terms that appeared more than 5 times are shown in Table I. From the set of these 37 terms 24 were finally selected as the contexts (see Table II).

The topic, objects and contexts were presented by the session organizer to the participants. The presentation should

Time	Family	Freedom
Traveling	Health	Enjoyment
Sport	Sleep	Success
Exercise	Music	Nutrition
Work	Pleasure	Sun
Friends	Satisfaction	Nature
Social interaction	Relaxation	Forest
Sharing	Harmony	Money

TABLE II
CONTEXTS FOR THE WELLBEING CASE, SELECTED FROM THE MOST COMMON TERMS GENERATED BY WORKSHOP PARTICIPANTS BY ASSOCIATION.

be conducted as neutrally as possible to avoid raising issues that refer to the value or opinion differences related to the topic. The presentation of the objects should be very plain so that no discussion is conducted related to them, i.e. basically they are just listed. On the other hand, the contexts are introduced with some detail. They are meant to serve as the common ground.

As a result of the preparatory step, the participants are aware of the contexts which are used in the analysis and should be ready to fill in a questionnaire that is presented to them in the next step.

1) Filling in the subjectivity tensor: The participants are asked to fill in a data matrix which typically consists of the objects as rows and the contexts as columns. Each individual's task is to determine how strongly an object is associated with a context. A graded scale, such as Likert from 1 to 5, can be considered beneficial.

There are several options regarding how the data collection can be conducted. It is possible to create a form on paper that is given to the participants to be filled in. Filling in the data takes place usually during the session because it is preceded by the introduction to the contexts. If there are any open questions related to the contexts, these are answered and discussed in a shared manner so that potential for creating a shared ground is maximized.

The data can also be collected with the help of some technological means. For instance, the participants may have access to a web page containing the input form, or the same functionality can be provided with mobile phone technology. In our wellbeing case, we used Google Docs to implement the questionnaire. This kind of web-based solution makes it easier to continue with the analysis as the data is readily in electronic form.

If the data has been gathered in paper form, there must be enough human resources available for typing these into the computer system. A simple solution is to have a spreadsheet file. In it, from each participant we now have a "data sheet". Together these sheets form the subjectivity tensor.

As a result of this step, the subjectivity data is ready to be analyzed. In this example, the tensor is of size $13 \times 4 \times 24$ ($\mathrm{subjects} \times \mathrm{objects} \times \mathrm{contexts}$), and each element takes a value between 1–5. The data analysis process is presented in the following.

2) Data analysis and visualization: The data collected in the previous task is analyzed using some suitable data analysis method. An essential aspect is to be able to present the rich data in a compact and understandable manner so that the conceptual differences are highlighted. In the following, we present an example where we look at some details of the tensor using histograms, and then form an overview using the Self-Organizing Map [15] algorithm. To show that the GICA method is independent from the choice of a specific methodology, we also apply multidimensional scaling (MDS) and neighborhood retrieval visualizer (NeRV) methods in the analysis.

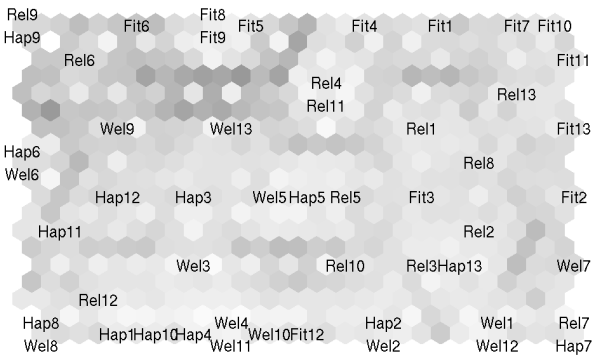

Fig. 3. Map of the subjects (numbered 1-13) and their views on wellbeing (Wel), happiness (Hap), fitness (Fit) and relaxation (Rel).

When the subject-object-context tensor is available, there are several options for analyzing it. The basic option is to consider all way-n matricizations These alternatives include creating a map of 1) the subjects and objects jointly based on the contexts (see Fig. 3), 2) the contexts based on how they were associated with the objects by each of the subjects (see Fig. 7), and 3) the subjects based on their responses considering the relationship between the objects and contexts. In this case, the subjects cannot be identified, neither are they grouped.

We also analyzed the combination of subjects and objects based on the context data using multidimensional scaling (MDS)[1] (see Fig. 4) and neighborhood retrieval visualizer (NeRV)[2] (see Fig. 5)[16]. The results obtained with MDS and NeRV confirm the SOM analysis results, especially concerning the status of fitness.

In the present analysis on the wellbeing concepts, one clear finding can be reported. Namely, a careful inspection of Fig. 3 reveals that the views on relaxation are widely scattered on the map whereas especially the concepts of happiness and fitness are much more concentrated on the map and therefore intersubjectively shared. Happiness becomes located on the left and lower parts of the map. Fitness is located on the upper and upper right parts. As a strikingly different case,

[1]The MDS analysis was conducted using the Matlab functions *pdist* and *cmdscale* with default parameters.

[2]The dredviz package used in the analysis with default parameters is available at http://research.ics.tkk.fi/mi/software/dredviz/.

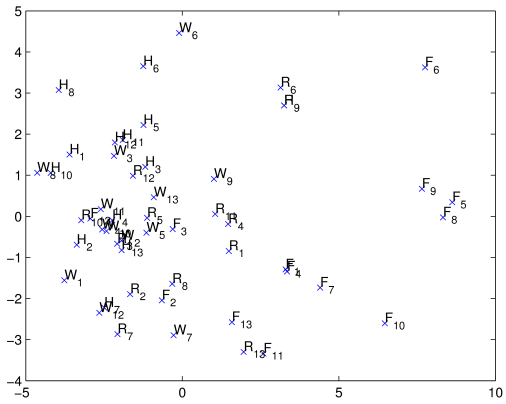

Fig. 4. Multidimensional scaling (MDS) analysis of the subjects (numbered 1-13) and their views on wellbeing (W), happiness (H), fitness (F) and relaxation (R).

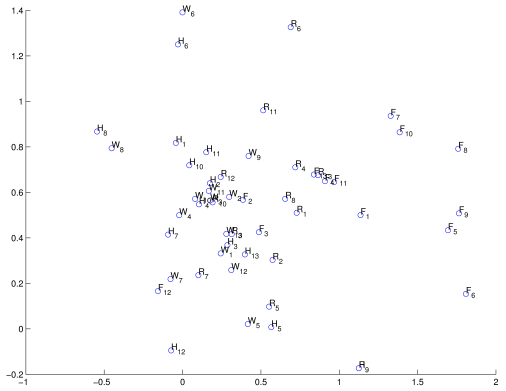

Fig. 5. Neighborhood Retrieval Visualizer (NeRV) analysis of the subjects (numbered 1-13) and their views on wellbeing (W), happiness (H), fitness (F) and relaxation (R).

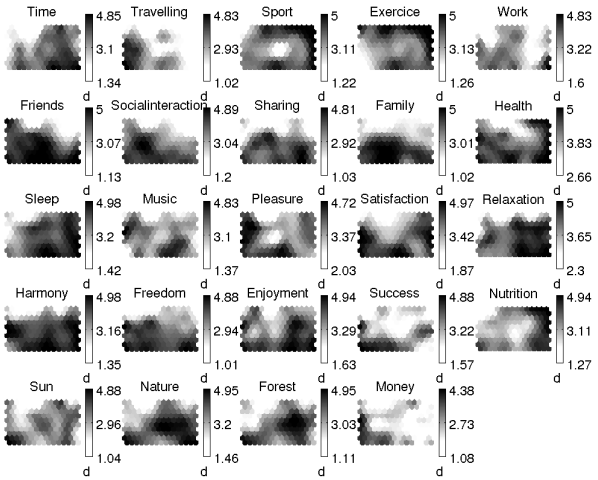

Fig. 6. Distributions of context items on the map shown in Figure 3. In these diagrams, a dark color corresponds to a high value (close to 5) and a light color to a low value (close to 1).

relaxation is not viewed in an uniform manner by the subjects. For example, for the subject 9, relaxation is located on the upper left corner of the map whereas the subject 7 is located on the opposite corner. The rest of the subjects are scattered around the map without any obvious pattern.

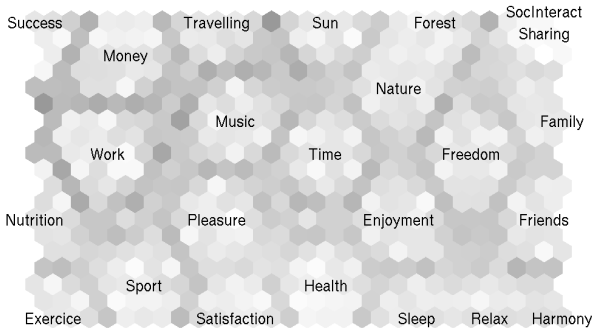

Fig. 7. Map of contexts.

In addition to considering the value of the context assessment of each subject-object pair shown in Fig. 3, one can also analyze the relationships between the distributions of each context item. The distributions are shown in Fig. 6. For instance, the distribution of the Exercise context on the map coincides very well with the object of Fitness in Fig. 3. The distribution of Exercise seems to be quite opposite to that of Traveling, Social interaction or Friends. This seems to indicate that the participants have viewed exercise to be separate from the social aspect of life. It is not a surprise that the distributions on pleasure and satisfaction coincide almost fully.

The relationships between the contexts can be made explicit by creating a map shown in Fig. 7. As an example of a clear result, one can pay attention to some specific pairs of contexts. Each item in the pairs "money-success", "sharing-social interaction", "sport-exercise" and "sleep-relaxation" can be found near one another on the map. They can therefore be considered as closely related contexts among the participants of this survey.

B. Text mining of subjectivity

Conducting a conceptual survey requires considerable amount of resources and therefore alternative means for obtaining subjectivity data are useful. In this section, we introduce the use of text mining in this task. The basic idea is to analyze a number of documents stemming from different persons and to compare the use of a set of words or phrases by them. The comparison is based on analyzing the contexts in which each person has used each word. The more similar the contextual patterns between two persons for a word, the closer the conceptions are considered to be. The accuracy of the result is, of course, dependent on how much relevant data is available.

1) Forming the subjectivity tensor: The GICA method based on text mining is illustrated by analyzing the State of the Union addresses by US presidents[3]. These speeches have been given since president George Washington in 1790. For the detailed analysis, we selected all speeches between 1980 and 2011 given by Jimmy Carter, Ronald Reagan, George Bush, Bill Clinton, George W. Bush and Barack Obama. The corpus has been used in many text mining studies (see e.g. [17], [18], [19]) but according to our knowledge the present analysis is unique in its kind.

JC	Jimmy Carter
RR	Ronald Reagan
GB	George Bush
BC	Bill Clinton
GWB	George W. Bush
BO	Barack Obama

TABLE III
PRESIDENTS AND THEIR ACRONYMS INCLUDED IN THE GICA ANALYSIS.

As in the previous case, the GICA method is applied by specifying a number of objects, contexts and subjects, and then filling in the data tensor. In this text mining case, populating the $O \times C$ matrix for each subject takes place by calculating the frequencies on how often a subject uses an object word in the context of a context word. A specific feature in this study was that each president has given the State of the Union Address several times. The basic approach would be to merge all the talks by a particular president together. However, a further extension was used, i.e., each year was considered separately so that each president is "split" into as many subjects as the number of talks he has given (e.g. Reagan_{1984}, Reagan_{1985}, ...). This is a sensible option because it also provides a chance to analyze the development of the conceptions over time.

In our study, the words "war", "peace", "business", "security", "progress", "justice", "freedom", "health", "education", "welfare", "community", "trust", "safety", "liberty" and "family" were selected as objects. The list of 559 context words was chosen to consist of meaningful words such as "action", "administration", "advantage", "aggression", "agriculture" "alien", "appreciation", "army" and "attack'[4]. Based on these lists and the original speeches, the $O \times C$ matrix was created in the following way. For each speech, the number of occurrences of any context word in the vicinity of each object word was counted. In our case, the vicinity was defined as 30 words preceding the object word. The contextual window cannot be the whole document because all the objects in a speech would obtain a similar status. On the other hand, a too short window would emphasize the syntactic role of the words. The limit 30 is admittedly arbitrary but according to unpublished experiments, varying the choice does not have a dramatic effect on the results. Another constant used in the experiment was the minimum number of occurrences of a object-context word pair. To be included in the analysis, a pair was required to occur more than 12 times. This choice was motivated by the analytical and visualization purposes.

[3]http://www.thisnation.com/library/sotu/

[4]The full list is too long to be shown here in its entirety but it can be downloaded at http://users.ics.tkk.fi/tho/sotuctx.txt

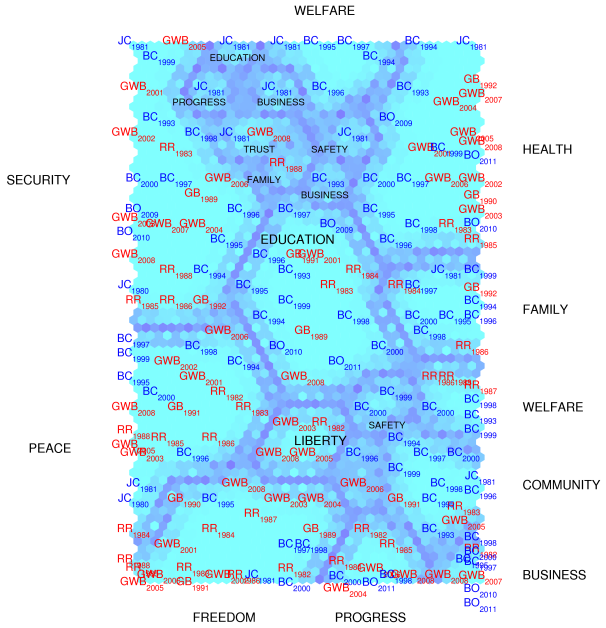

Fig. 8. SOM-based GICA map of object words from speeches by US presidents between 1980 and 2011.

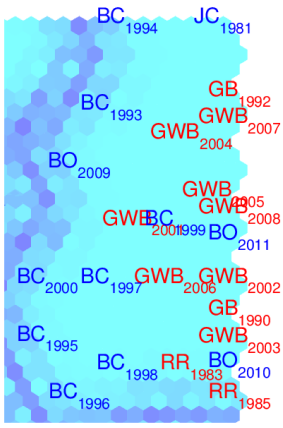

Fig. 9. A zoomed view into the health area of the GICA map.

2) Visualization: We wished to be able to show all the data in the same visualization, shown in Fig. 8. The parameter choices lead into the selection of 180 occurrences which consist of subject-object pairs. An additional feature here is that there may be several instances of these pairs for different years. The labels on the map consist of the acronym of each president and the year when the State of the Union address was given:

Fig. 9 shows a detailed view on the health area of the overall map shown in Fig. 8. Two specific conclusions can be made. First, a general tendency is that the handling of the health theme forms two clusters, the democrats of the left and the republicans on the right. However, the second conclusion is that in Barack Obama's speeches in 2010 and 2011, he has used the term in a way that resembles the republican usage.

We also created a map of people, i.e., a map of US presidents (Fig. 10). This analysis is based on a flattening of the data tensor through way-1 matricization so that speeches form the rows of the matrix and the columns correspond to the occurrences of object words associated with the context words.

IV. Discussion

We have introduced 1) a novel multi-disciplinary theoretical framework in which a class of problems related to human communication can be made explicit and 2) a qualitative-quantitative methodology as a practical solution for these kinds of problems. The main idea is to provide computational methodology that can be used to represent, analyze and visualize situations in which different people have varying underlying conceptions that may hinder successful communication.

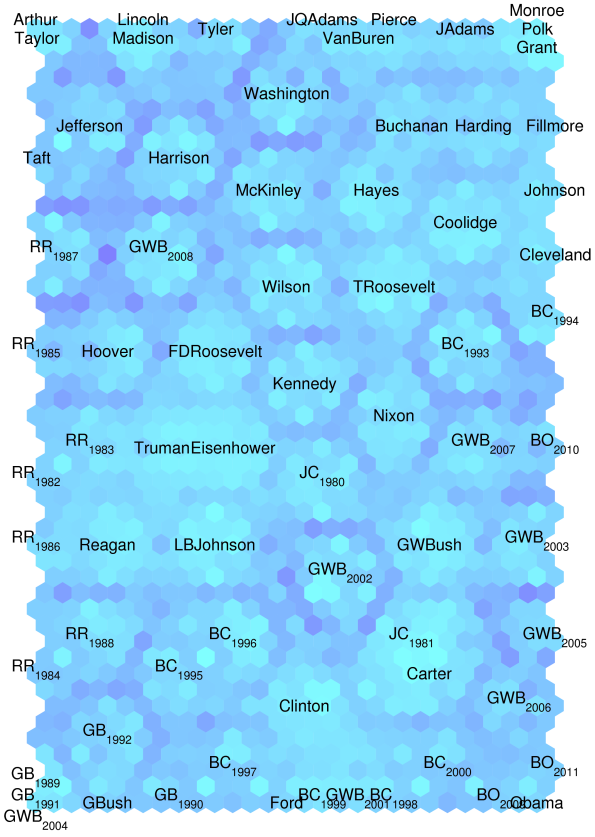

Fig. 10. SOM-based GICA map of all US presidents based on their State of the Union addresses. The map includes all presidents and additionally a positioning of the yearly speeches between 1980 and 2011.

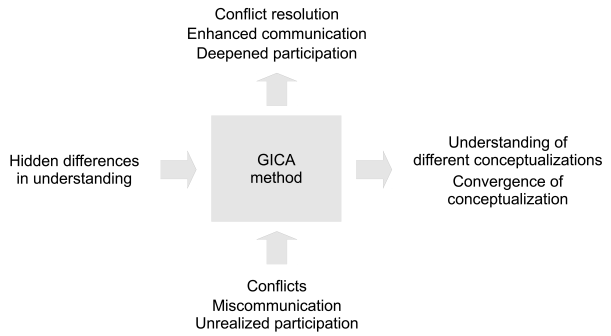

Fig. 11. Potential uses and application areas of the GICA method.

Future research includes the further development and application of three-way analysis methods in GICA. As an empirical research question, we plan to study in relation to different tasks and domains whether the GICA method helps, and to which extent, in improving communication over time. The tasks and domains may include, for instance, environmental decision making, crisis management, interprofessional and interdisciplinary communication and conceptual change processes (see Fig. 11, cf. also [20]).

One potentially important application area for the GICA method is semantic web. In semantic web, knowledge representations are built with some kind of objectivity or intersubjectivity in mind. However, in practice the representations created by different people and organizations even in a narrowly defined domain tend to vary. Various semantic mapping techniques have been developed but, in general, it seems necessary to ground semantic representations in their relevant contexts [21], [22]. The GICA method provides a framework in which both grounding in context and modeling subjective or organizational semantic variation can take place.

ACKNOWLEDGMENTS

The authors wish to thank Academy of Finland and the Department of Information and Computer Science, Aalto University School of Science for the funding of this research. We are also grateful to the EIT ICT Labs and the participants of the Wellbeing Innovation Camp 2010 with whom the data for the first case study was gathered. The first author (T.H.) also wishes to thank Lund University Cognitive Science and prof. Peter Gärdenfors who hosted a visit in 2003 during which the first GICA-like experiment took place, and the EU Commission for the funding of the META-NET Network of Excellence and the MultilingualWeb project within which many useful discussions have taken place related to the practical needs as well as current limitations concerning semantic representations.

REFERENCES

[1] T. Honkela and A. M. Vepsäläinen, "Interpreting imprecise expressions: Experiments with Kohonen's self-organizing maps and associative memory," in *Artificial Neural Networks*, vol. I. Amsterdam, Netherlands: North-Holland, 1991, pp. 897–902.

[2] T. Honkela, "Self-Organizing Maps in Natural Language Processing," Ph.D. dissertation, Neural Networks Research Centre, Helsinki University of Technology, Espoo, Finland, 1997.

[3] T. Honkela, V. Könönen, T. Lindh-Knuutila, and M.-S. Paukkeri, "Simulating processes of concept formation and communication," *Journal of Economic Methodology*, vol. 15, no. 3, pp. 245–259, Sep. 2008.

[4] N. Janasik, T. Honkela, and H. Bruun, "Text mining in qualitative research: Application of an unsupervised learning method," *Organizational Research Methods*, vol. 12, no. 3, pp. 436–460, 2009.

[5] B. Glaser and A. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Publishing Company, 1967.

[6] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335–346, 1990.

[7] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[8] B. Berlin and P. Kay, *Basic Color Terms*. University of California Press, 1969, vol. 37.

[9] T. Regier, P. Kay, and R. S. Cook, "Focal colors are universal after all," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 23, pp. 8386–8391, 2005.

[10] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning subjective language," *Computational Linguistics*, vol. 30, no. 3, pp. 277–308, January 2004.

[11] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation," in *EMNLP*, 2009, pp. 190–199.

[12] A. Mustajoki, "Modelling of (mis)communication," in *Prikladna lingvistika ta ligvistitshni tehnologii: Megaling-2007*, 2008, pp. 250–267.

[13] ——, "A multidisciplinary and multidimensional approach to risks and causes of miscommunication," *Language and Dialogue*, in print.

[14] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[15] T. Kohonen, *Self-Organizing Maps*, ser. Springer Series in Information Sciences. Berlin, Heidelberg: Springer, 2001, vol. 30.

[16] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *Journal of Machine Learning Research*, vol. 11, pp. 451–490, 2010.

[17] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2006, pp. 424–433.

[18] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant, "Discovering interesting usage patterns in text collections: integrating text mining with visualization," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ser. CIKM '07. New York, NY, USA: ACM, 2007, pp. 213–222.

[19] A. A. Balinsky, H. Y. Balinsky, and S. J. Simske, "On Helmholtz's principle for documents processing," in *Proceedings of the 10th ACM symposium on Document engineering*, ser. DocEng '10. New York, NY, USA: ACM, 2010, pp. 283–286.

[20] T. Honkela, "Von Foerster meets Kohonen - approaches to artificial intelligence, cognitive science and information systems development," *Kybernetes*, vol. 31, no. 1/2, pp. 40–53, 2005.

[21] T. Honkela and M. Pöllä, "Concept mining with Self-Organizing Maps for the Semantic Web," in *Proceedings of WSOM'09*. Springer, 2009, pp. 98–106.

[22] K. Hakkarainen, R. Engeström, S. Paavola, P. Pohjola, and T. Honkela, "Knowledge practices, epistemic technologies, and pragmatic web," in *Proceedings of I-KNOW'09 and I-SEMANTICS'09: the 4th AIS SigPrag International Pragmatic Web Conference Track (ICPW 2009)*. Verlag der Technischen Universität Graz, 2009, pp. 683–694.