Directions for E-science and Science 2.0 in Human and Social Sciences

Timo Honkela *

Aalto University School of Science and Technology
Department of Information and Computer Science
P.O.Box 15400
00076 Aalto, Finland

Abstract.

In this review and tutorial article, new developments towards extended use of information and communications technologies in science are discussed. The focus is in human and social sciences, specifically in linguistics and economics. Some challenging epistemological issues are handled in detail including the subjective and intersubjective nature of human knowing and how it influences scientific practices. Examples related to the use of data and text mining in human and social sciences are provided. Also the use of social simulation is considered in some detail. The article is concluded by a discussion on some potential practical implications for future scientific practices.

Dans cet aperçu, et ce tutorial, de nouveaux développements pour l'utilisation étendue des technologies d'information et de communication dans les sciences sont discutées. L'accent est mis les sciences humaines et sociales, en particulier la linguistique et l'économie. Certains défis épistémologiques sont traités dans le détail, y compris le caractère subjectif et intersubjectif de la connaissance humaine et comment elle influence les pratiques scientifiques. Exemples liés à l'utilisation des données et de text mining dans les sciences humaines et sociales sont fournis. De plus, l'utilisation de la simulation sociale est considérée en détail. Une discussion sur un certain nombre d'implications pratiques possibles pour l'avenir de pratiques scientifiques conclu l'article.

1 Introduction

Scientific practices are in flux due to the developments in information and communication technologies. Not only can we easily find and send scientific information between researchers, but we can also co-construct knowledge with the help of increasing number of tools meant for this purpose. The developments to this direction have been coined with the term Science 2.0. Moreover, computational modeling and simulation have become more and more commonplace in almost all scientific disciplines. These developments are next considered in some detail.

*A keynote talk in MASHS 2010, Computational Methods for Modelling and Learning in Social and Human Sciences, Lille, France, 10-11 June 2010. Appeared in the MASHS 2010 conference proceedings, pp. 119-134, edited by C. Biernacki, E. Masson, A. Lendasse and E. Séverin, published by Multiprint.

1.1 Science 2.0

Science 2.0 builds on the technologies of Web 2.0. Blogs, wikis and other social sharing and interaction tools allow scientists to interact and make their data and interpretations available for others in novel ways[1]. In Science 2.0, in addition to research articles, also research data, implemented methods, detailed results and comments are available online. Anyone can review the data, analyses, theories, interpretation and conceptual frameworks and use them in further experiments and theory construction. Open data repositories allow the data to be aggregated and new conclusions and interpretations to be found.

Science 2.0 continues and extends the tradition of publishing open source software[2] and open access publishing of scientific articles. In a recent survey, 39 articles were found in which an open access citation advantage was shown and 7 in which no advantage or ascribing the advantage to factors unrelated to open access publication [1]. These studies typically show a 25-250% open access citation advantage or more, but, according to Wagner, the higher end of the range may prove illusionary [1]. Thus, one reason for open access publishing is that it can be a major incentive for scholars to be cited. However, there are even more important reasons for sharing the data and methods behind published scientific research results. For instance, the aspects of quality assurance through reproducibility and efficiency through decreasing amount of overlapping work and increasing bandwidth of idea-level and implementation-level communication can be mentioned.

In general, research should be reproducible. However, much scientific research is too complicated and the published methods and data are not detailed enough for other scientist to reproduce them. The lack of detail may be due to the page limits of journals On the other hand, scientific papers nowadays also point to supplementary materials on the internet. Especially in in mathematical and computing sciences, reproducibility is possible and there are efforts towards increasing attention to this[3].

1.2 Computational science

Where Science 2.0 refers to new practices in conducting science with the help of communications and collaborations technologies, *computational science* builds on modeling and simulation of real world or anticipated phenomena based on massive data sets. Traditionally, this field has been dominated by applications related to natural sciences and engineering, but also human and social sciences have started to use computational models as a research tool. This is reflected

[1]This paragraph is based on the Wikipedia article on Science 2.0 (`http://en.wikipedia.org/wiki/Science_2.0`)

[2]Like *SOM Toolbox* for Matlab (`http://www.cis.hut.fi/somtoolbox/`), *FastICA* for Matlab (`http://www.cis.hut.fi/projects/ica/fastica/`), *dredviz* software package for dimensionality reduction in information visualization (`http://www.cis.hut.fi/projects/mi/software/dredviz/`), and *Morfessor* software (`http://www.cis.hut.fi/projects/morpho/`) as notable local examples.

[3]Consider, e.g., `http://reproducibleresearch.net/`

by the increasing number of research papers published and scientific conferences devoted to this area, the MASHS series as a notable example[4].

Reproducibility may be essential in the area of computational science. According to practical experience, the peer review process in computational science generally does not necessarily provide as effective a filter as it does for experiment or theory [2]. Related to a computational science paper, the code may have hidden defects, it might be applying algorithms improperly, or its spatial or temporal resolution might be inappropriately coarse [2].

1.3 Computational linguistics

Computational linguistics is an area in which computers have been used for a relatively long time as a research tool. Linguistics can be considered to particularly interesting from the point of view of scientific practice and scientific representation because language is a central means for representing and communicating scientific results. In linguistics, the representational levels are intertwined, as illustrated in Fig. 1. The community of language users (basically all human beings and in some sense also an increasing number of computational artefacts that process natural languages such as machine translation tools) produce and create language. By producing, we refer to the generation of linguistic expressions according to existing "rules and principles" including syntactic rules and lexical items. Creation refers to the fact that these abovementioned rules and principles are occasionally reformulated in their details by introducing new words to a language or by promoting new constructions that are gradually taken into common use.

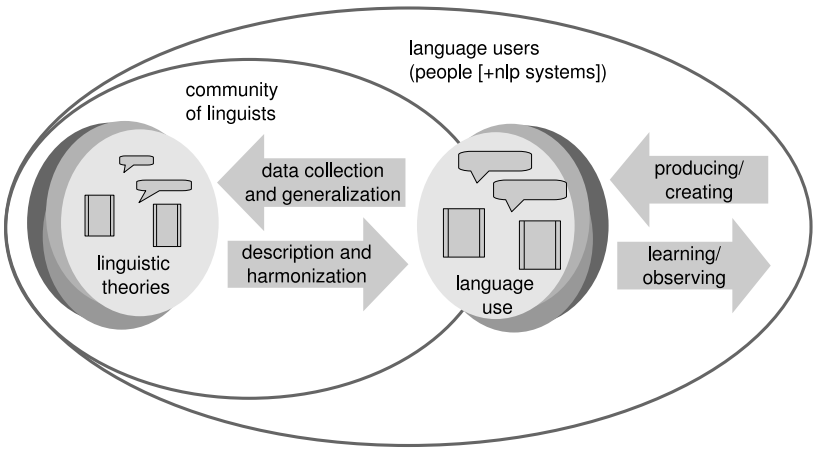

Fig. 1: Social construction of language and linguistic knowledge.

Vygotsky observed how higher mental functions have developed in cultural

[4]For MASHS 2008 and 2010 web sites, see `http://mashs08.univ-paris12.fr/` and `http://mashs2010.free.fr/`.

groups and individually through social interactions [3]. Through these interactions, children learn the habits of their culture, including speech patterns, written language, and other symbolic knowledge. The specific knowledge gained by children represents the shared knowledge of a culture including the norms related to language use. This process is known as internalization and the opposite process as externalization [3]. Linguistic norms are not static but they emerge, evolve, and disintegrate at a sociocultural level [4]. They are implicitly represented in linguistic expressions and explicitly represented as externalized rules. The explicit representation as rules is one of linguists' activities as illustrated on the left hand side of Fig. 1.

From the methodological point view, a distinction into two basic paradigms in computational linguistics can be made. The first paradigm relies on explicit encoding of linguistic knowledge by linguists based on their intuitions, potentially supported by corpus linguistic studies (see, e.g., [5, 6]). The second paradigm is based on creating language models using statistical (machine learning) methods (see, e.g., [7, 8, 9]). The latter paradigm relies on the availability of large corpora that can be used to train the models. In addition to the methodological differences, there are epistemological issues that are discussed in Section 3.

1.4 Computational economics

Computational economics is a research area in the intersection of economics and computer science. Specific areas include, for instance, computational econometrics, computational finance, computational modeling of macroeconomic systems, and agent-based economic modeling [10]. Many applications of neural networks and statistical machine learning exist (see, e.g., [11, 12, 13]).

In the following, a specific topic that form a link between computational economics and computational linguistics is discussed. The traditional notion of uncertainty in decision making does not cover the uncertainties caused by differences in conceptual systems of individual agents within a community. We claim that in all transactions including symbolic/linguistic communication the differences in the underlying conceptual systems play an important role [14]. For instance, serious efforts have been made to harmonize or to standardize the classification systems used by business agents, e.g., using Semantic Web technologies [15]. However, even if the standardization is conducted, there can not be any true guarantee that all the participating agents would share the meaning of all the expressions used in the business transactions in various contexts [14].

We can consider an example in which a buyer agent expresses the wish for finding an item belonging to some category. It may very well be that the selling agent understands the query differently and therefore the items considered as candidates by the selling agent differs from the buyers intentions. Within traditional epistemological theory, this situation could be considered within set theory: by a symbol the agents refer to different sets of items. However, as many features (or quality dimensions) are continuous, it appears to be more natural to consider the situation within a continuous multidimensional space. Moreover, the exact features of an item may not be known but they need to be considered

as a probability distribution. An implication is that in business transactions there should be means for checking what is a meant by some expressions via an access to a broader context (cf. symbol grounding). Moreover, rather than relying solely on a standardized conceptual system, one could introduce mechanisms of meaning negotiation. Before two business agents get into negotiation about, for instance, the price of some commodity, they should first check if they agree on what they refer to by the expressions that are used in the negotiation. This concern is valid both for human and computerized systems, even though humans are usually capable of conducting meaning negotiations even when they are not aware of it. [14]

2 Data and text mining in human and social sciences

In the following, some examples of using data and text mining in the area of human and social sciences are considered.

2.1 Data mining in political sciences

The complex phenomena of political science are typically studied using qualitative approach, potentially supported by hypothesis- driven statistical analysis of some numerical data. We have examined the use of the self-organizing map in this area and explored the relationship between parliamentary election results and socio-economic situation in Finland between 1954 and 2003 [16]. The variable maps (or component planes, as they are traditionally called) show the distribution of each separate variable on the map. These are presented in Fig. 2. A change that took place in the late 1970s is clearly discernable. Many dependences between variables changed their features. Correlations turned from negative to positive and vice versa. More details and a qualitative analysis of the results are available in [16].

2.2 Text mining and qualitative research

Text mining using the self-organizing map (SOM) [17] presents an interesting methodological opportunity for qualitative research. Qualitative researchers aim to gather rich understanding of human behavior and the reasons for the behavior. In qualitative research, small but focused samples are therefore more often used, rather than large samples. We have argued that the SOM is particularly efficient in improving inference quality within qualitative research, with regard to both confirmatory and exploratory research [18]. Within the theory-driven or deductive mode of qualitative research, the SOM can be used to test the adequacy of conceptual frameworks created before the analysis of the data. In the data-driven or inductive mode, the SOM can be applied in creating emerging category systems describing and explaining the data [18]. By qualitative research we here mean the analysis of interviews and similar kind of free form language data, rather than, e.g., the analysis of qualitative variables (see, e.g., [19].

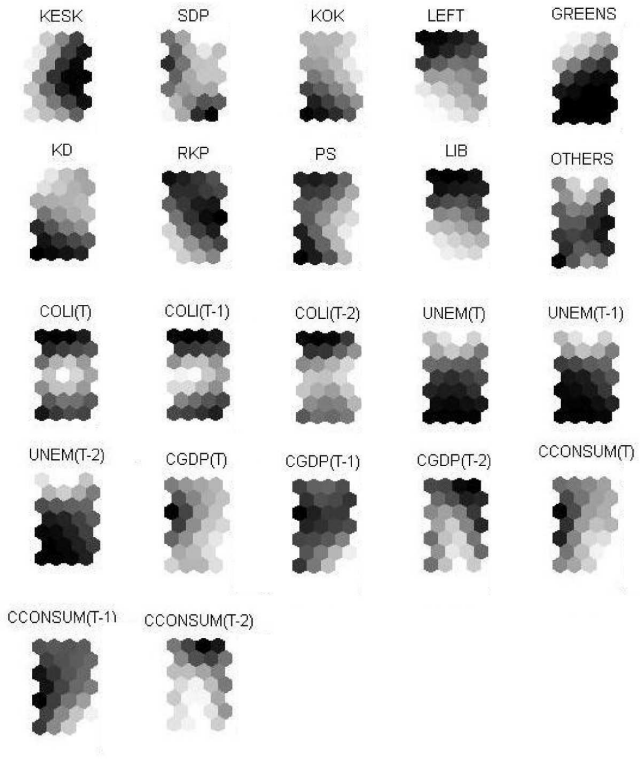

Fig. 2: Some variable maps of variables used in the study. The acronyms for the political parties are as follows: KESK: Centre Party of Finland, SDP: Social Democratic Party of Finland, KOK: National Coalition Party, VAS: Left Alliance, GREENS: Green League, KD: Christian Democrats in Finland, RKP: Swedish People's Party, PS: True Finns, and LIB: Liberals. National economic conditions are analyzed using four measurements: Change of Cost of Living Index (COLI), Unemployment Rate (UNEM), Change of Gross Domestic Product per Capita (CGDP), and Change of Total Consumption per Capita (CCONSUM). These four monetary values are transformed into constant prices of the year 2000. For each measurement, there are three variables included in the data: the first at elections year (marked with COLI(T), UNEM(T), CGDP(T), and CCONSUM(T)), the second at a year before elections (marked with COLI(T-1), etc.) and the third at two years before elections (marked with COLI(T-2), etc.)

The SOM (and related methods) can be considered as a quantitative method or research tool that is particularly well suited to the aim of respecting complexity rather than trying to do away with it. The SOM can produce not only one but a multitude of perspectives on some data. In relation to very large data sets

of the kind, some of these multiple perspectives might be such that no human would, even in principle, be able to produce them. This follows from the fact that the computational method can be used to process writings or sayings of thousands or even millions of persons, something that is beyond the scope of any individual researcher. Yet applying the SOM allows us access to potentially highly relevant and novel categories and patterns that "really are there," even if we do not as yet know it. This would appear to be particularly true when it comes to various nonconscious categorizations. Thus, it would appear that applying the quantitative method of the SOM could take us even beyond situational analysis in that it is capable of revealing subconscious operations of the human mind, which the consciously operating human mind of the situational analyst will never be able to discover [18]. In general, a cartographer of social life can greatly benefit from taking the text mining results into account.

2.3 Science mapping

In the following, we exemplify the potential of using SOM through a case study in which the amount of qualitative data was a central factor. In 2004, the Academy of Finland, one of the country's largest funding agencies, commissioned a study to investigate to what extent and how the academy had promoted interdisciplinary research in its funding and to recommend how the academy could improve its capabilities in fostering interdisciplinary research [20]. Bruun and his colleagues used applications to the Academy of Finland as empirical material, classifying the applications on the basis of a qualitative analysis of their contents. They found that more than 40of a sample of 324 successful research applications proposed to do interdisciplinary research [21]. During the analysis, 266 applications were read and carefully analyzed and qualitatively assessed. This process took one researcher approximately 5 to 6 weeks.[18].

As a continuation of this manually conducted qualitative analysis [21], the Academy of Finland in 2006 commissioned another study to investigate whether text mining based on the SOM could be used to support assessment of the applications. At the Adaptive Information Research Centre, a collection of 3,224 applications was analyzed by Timo Honkela and Mikaela Klami. A collection of 1,331 term candidates was automatically extracted using a reference corpus-based method [22]. The method is based on the following idea: Words and phrases that are relatively more common in the text collection under study than the reference corpus are good term candidates [22]. Term candidates that belonged to categories such as names of persons, organizations, or places were then manually left out. In the end, there was a collection of 1,200 terms. The 3,224 application documents were encoded as term distribution patterns. The SOM algorithm organized the documents into a map in which similar applications are close to each other and in which thematic areas emerged (see Fig. 3).

One interesting finding was related to the division into research councils. The Academy of Finland organizes its activities into four councils: (a) health, (b) biosciences and environment, (c) culture and society, and (d) natural sciences and engineering. In the SOM analysis, the applications were distributed on the

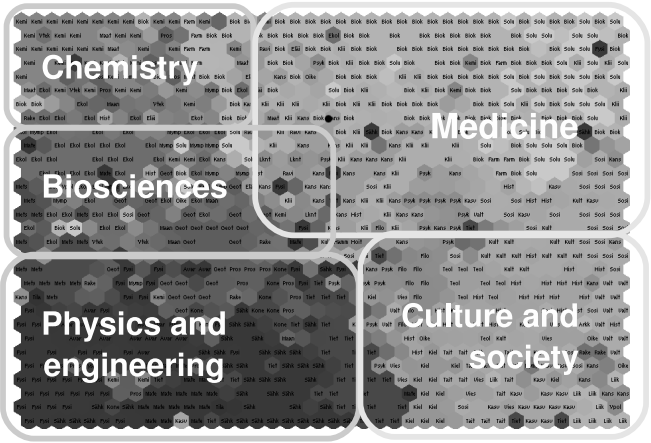

Fig. 3: Map of Finnish science.

map in a manner that mostly followed the division into the councils, with one important exception: the natural sciences and engineering research council was split into two parts, biosciences and environment. Specifically, the research related to chemistry within natural sciences and engineering was clearly separated from the research on, for instance, physics and engineering sciences. Moreover, research on chemistry was also closer to the area occupied by the health research council than the other disciplines in natural sciences and engineering.

3 Epistemological considerations

Next, some epistemological issues are considered in some detail.

3.1 Explicit logical formalization of knowing

One underlying motivation of this article is the recognition of a potential need to increase the variety of methods that are used to deal with issues within philosophy of language. To express the basic situation in a slightly simplified manner, the formal methodological realm of philosophy of language is still largely dominated by predicate logic. Many philosophers, including Gottlob Frege, Alfred Tarski and Rudolf Carnap, have been more or less skeptical about formalizing natural languages, but many of them have relied on a certain level of formalization. Some of the prominent members of this tradition of formal semantics include Alfred Tarski, Rudolf Carnap, Richard Montague and Donald Davidson. Recent works on formal semantics have been conducted, for instance, by Jon Barwise, Robin Cooper and Barbara Partee [23]. The works on analytical philosophy of language tend to focus on some particular aspects of natural language such as truth conditions, and the role of quantifiers and connectives. Maybe the most striking example of formalization of natural language is the

work of Richard Montague. Montague's thesis was that there is no essential difference between the semantics of natural languages (like English) and formal languages (like predicate logic), i.e., there is a rigorous way how to translate English sentences into an artificial logical language [24]. Montague grammar is an attempt to link directly the syntactic and semantic level of language. In order to do so, Montague defined the syntax of declarative sentences as tree structures and created an interpretation of those structures using an intensional logic. The end result was a focus on such aspects of language that nicely fit with the theoretical framework. Examples of language considered includes sentences like "Bill walks", "every man walks", "the man walks", and "John finds an unicorn" [24]. It may be fair to say that most of the linguistic phenomena are set aside. Montague even assumes that the original sentences can be considered unambiguous even though ambiguity is a central phenomenon in language at many levels of abstraction. The idea of being rigorous may be considered a proper stand but it often leads to the negligence of the original complexity of the phenomenon being considered [25].

Many philosophers outside the analytical tradition have already for some time criticized the approach of logical formalization within philosophy of language. For instance, representatives of phenomenology (e.g. Edmund Husserl and Martin Heidegger), hermeneutics (e.g. Martin Heidegger and Hans-Georg Gadamer) and critical theory (e.g. Max Horkheimer and Jürgen Habermas) have presented alternative views. Richard Rorty [26] attacks the correspondence theory of truth (that truth is established by directly comparing what a sentence asserts to the "facts applying), and even denies that there are any ultimate foundations for knowledge at all. He calls for a socially-based theory of understanding. He also strongly criticizes the notion of truth: Truth is not a common property of true statements, and the good is what proves itself to be so in practice. Rorty combines pragmatism (cf. e.g. John Dewey and Charles S. Peirce) with the philosophy of language by later Wittgenstein which declares that meaning is a social-linguistic product. It is far from obvious that communication between speakers of one and same language would be based on commonly shared meanings as often suggested by the proponents of formal semantics, either explicitly or implicitly. This leads to the rejection of the idea of an idealized language user and to the rejection of the possibility to consider central epistemological questions and natural language semantics without considering subjectivity and variability. In other words, the language of a person is idiosyncratic and based on the subjective experiences of the individual. For instance, two persons may have a different conceptual or terminological density of the topic under consideration. A layperson, for instance, is likely to describe a phenomenon in general terms whereas an expert uses more specific terms. Moore and Carling [27] state that languages are in some respect like maps. If each of us sees the world from our particular perspective, then an individual's language is, in a sense, like a map of their world. Trying to understand another person is like trying to read the map of the other, a map of the world from another perspective [27].

Rather than using first-order predicate logic, modal logic and other simi-

lar formal languages as a basis for theory formation within epistemology, it is strongly suggested that they might even be mostly replaced by probability theory, matrix algebra, dynamical systems theory and other statistical and mathematical methods that seem to be better suited for building epistemological theories in order to be able to deal with continuous, multidimensional and dynamical phenomena that are inherent in knowledge formation and natural language understanding.

Notions such as a symbol or a proposition may still be useful in some theoretical contexts but they may rather be seen as abstractions that are emergent outcomes of some highly complex processes. However, even in the context of philosophy of language, it might be less misleading to use terms 'word' and 'sentence' rather than 'symbol' and 'proposition' and discuss the emergence of symbol-like and proposition-like phenomena through theoretical tools that are suited to capture the nature of those emergent processes. Following the criticism by Richard Rorty towards the notion of truth, discussed earlier in this article, one may really question the usefulness of the notions of truth as a useful building block in realistic epistemological theories. Truthlikeness already seems to capture better the nature of 'sentential knowledge' but even that term might draw the attention away from the idea of language being primarily a tool for communication of various kinds. Moreover, human understanding of the world and of the relationship between language use and perception and action within the world is based on a long learning process for which the genotype gives a certain basis but which is mainly determined by the individual interaction with the world including other human beings and the social and cultural context. In general, the centrality of learning processes emphasizes the need to consider the statistical aspects related to learning and using language. Language learning seems to be essentially a statistical process. There are some researchers such as Jerry Fodor who suggest that linguistic skills and even conceptual contents are innate in the mind. However, it seems that the arguments supporting the centrality of learning proposed, e.g, by Patricia Churchland, Paul Churchland, Andy Clark and Paul Smolensky are more realistic. This line of thought leads directly to the use of neural network models in modeling processes of language learning, understanding and generation. In language use, many specific situations can be readily analyzed as probabilistic questions.

3.2 Challenge of translation

Quine [28] presents a situation in which one is confronted with a situation in which one must attempt to make sense of the utterances and gestures that the members of a previously unknown tribe make. Quine claimed that it is impossible, in such a situation, to be absolutely certain of the meaning that a speaker of the tribe's language attaches to an utterance. For example, if a speaker sees a rabbit and says "gavagai", is she referring to the whole rabbit, to a specific part of the rabbit, or to a temporal aspect related to the rabbit. If one considers the point of view of radical constructivism [29, 25] and the symbol grounding problem [30], there can practically even be an infinite number

of conceptualizations of the situation. Maybe the members of the tribe not only consider the whole rabbit or some parts or aspects of it as potentially relevant points of reference but, e.g., due to their cultural context they consider some other patterns of perception. Namely, considering the complex pattern recognition process, it is far from trivial to create a perception of a rabbit from the raw visual and auditory input. Quine [28] mentions that one can form manuals of translation. The observer examines the utterances as parts of the overall linguistic behavior of the individual, and then uses these observations to interpret the meaning of all other utterances.

Quine continues that there will be many such manuals of translation since the reference relationship is indeterminate. He allows that simplicity considerations not only can be used to choose between competing manuals of translation but that there is even a remote possibility of getting rid of all but one manual. It seems that propositional logic as the underlying epistemological framework unnecessarily complicates the consideration. For Quine it was necessary to consider a number of logically distinct manual of translation hypotheses. However, if one considers the issue within the framework of statistics, probability theory and continuous multidimensional representations of knowledge (such as conceptual spaces [31]), one can consider the conditional probability of different hypotheses and partial solutions which do not need to be logically coherent. Moreover, the search for translation mappings can be seen as a process that may (or may not) converge over time. For Quine meaning is not something that is associated with a single word or sentence, but is rather something that can only be attributed to a whole language. The resulting view is called semantic holism. In a similar fashion, the self-organizing map specifies a holistic conceptual space. The meaning of a word is not based on some definition but is the emergent result of a number of encounters in which a word is perceived or used in some context. Moreover, the emergent prototypes on the map are not isolated instances but they influence each other in the adaptive formation process.

3.3 Implications for scientific practices

The impact of categorical thinking can be widely seen in scientific practices and in the society, in general. Conceptualization, categorization and search for regularities are, of course, essential tools of science, but they also have a downside. Perceptual psychology talks about a phenomenon known as categorical perception. Humans easily interpret a perception as belonging to some category, even though they have been picked between the categories. When a perceived phenomenon is categorized, it is even too easy to apply related rules to this specific instance. The idea of rules as seen in natural sciences, does not necessarily suffice when complex biological, human and social phenomena are considered. Von Foerster has even stated that if science is about finding universal laws or rules of the form "x follows from y", the majority of the phenomena in the world are uninteresting. This, in itself, is probably only one perspective into the well-known tension between natural sciences and the humanities.

For all fields of science, the use of natural language is in common. Natural

anguage expressions are often ambiguous or imprecise, but the transmission of precise meanings is, of course, the basic goal of scientific practices. However, the interpretation of expressions can not be completely harmonized by the scientific community either. Each individual has a rich experience about the world. The interpretation of a specific word or phrase may not be exactly the same as for another person. Learning processes in the human brain seems to give rises to unique patterns and models of statistical nature. These patterns reflect the general socio-cultural level, but they are not in a simple way shared between people. Although each word is given a careful definition, all people interpret the word through their own unique experience of the world. An antidote against this subjectivity in each field of science is the education and other related practices designed to ensure that the scientists use the key terminology in a sufficiently similar way. This is probably easiest to ensure when there are terms which scientists in other fields, let alone laymen are not even expected to be familiar with.

4 Socio-cognitive modeling

Socio-cognitive modeling is a new research area that merges aspects of computer science, social sciences and cognitive science. The basic idea is to model interlinked social and cognitive phenomena. Cognition and intelligent activity are not only individual processes but ones which rely on socio-culturally developed cognitive tools. These include physical and conceptual artifacts as well as socially distributed and shared processes of intelligent activity embedded in complex social and cultural environments [32].

4.1 Social level of reality

At the socio-cultural level, humans create and share conceptual artifacts such as symbols, words and texts. These are used as mediators between different minds. In communicating and sharing knowledge, individuals have to make a transformation between their internal representation into an explicit representation to be communicated and vice versa, as Vygotsky pointed out already in the 1930s. The internalization and externalization processes take place as a continuous activity. In externalization, the internal view is externalized as explicit and shared representations. Vygotsky also investigated child development and how this was guided by the role of culture and interpersonal communication [3]. He observed how higher mental functions develop historically in cultural groups and individually through social interactions. The specific knowledge gained by children represents the shared knowledge of a culture including the social norms, e.g., related to language use. In our research, we are interested how norms emerge, evolve, and disintegrate at a sociocultural level, how the norms are internalized and externalized by individuals, how they are followed or occasionally deliberately not followed, and how they are implicitly represented in linguistic expressions and explicitly represented as externalized rules.

4.2 Social simulation

One approach in socio-cognitive modeling is social simulation. It aims at exploring and understanding of social processes by means of computer simulation. Social simulation methods can be used to to support the objective of building a bridge between the qualitative and descriptive approaches used in the social sciences and the quantitative and formal approaches used in the natural sciences. Collections of agents and their interactions are simulated as complex non-linear systems, which are difficult to study in closed form with classical mathematical equation-based models. Social simulation research builds on the distributed AI and multi-agent system research with a specific interest of linking the two areas. The research area of simulating social phenomena is growing steadily (see, e.g., [33]).

Schwenk and Reimer have built a social simulation model to study the processes of social influence [34] that partly builds on the research on heuristics [35]. They examined the interaction of decision strategies and features of the communication network. Schwenk and Reimer's simulation model was contextualized by a scenario which they adapted from Lazega [36]. In this scenario, a group of lawyers who are partners in a law firm gather in a meeting in order to decide about topics concerning the firm, for instance, the branch of business in which the firm should further expand [34]. In the simulation model, the lawyers were represented by a set of 21 agents, each having a certain preference for a branch of business into which the firm should expand. In more detail, each agent l_i was associated with both a value d_i of a decision variable D, which contains the discrete decision alternatives, and a value w i of an individual status variable W. A directed graph G, described a network of directed communication channels c_{ji} between the agents L. Each agent l_i is assigned a decision procedure f out of a set of decision procedures F. This function f consisted of a contact rule r_c and a decision rule r_d and maps an agent's actual decision state d_{j_n} onto its subsequent state d_{j_n+1}. the dynamic evolution of the model was then based on the iterated and sequential call of this decision rule f.[34] The result of the simulation was such that the impact of the agents' decision strategies on the dynamics as well as the outcomes of the influence process depended on features of their social environment. This behavior particularly clear when the agents contacted all of the neighbors with whom they were connected [34]. From our point of view, an interesting extension of this work would be to combine the influence process modeling with some more specific consideration of the conceptual models of the agents.

Nishizaki, Katagiri and Oyama have developed an agent-based simulation model in which artificial adaptive agents have mechanisms of decision making and learning based on neural networks and genetic algorithms [37]. They compare the results of their simulation analysis with the ones of a mathematical model related to the potential occurrence of strikes in a labor market. One result stemming from the simulation model was that individuals behaved cooperatively and that the prisoners' dilemma could be escaped. The earlier model was based on rationality (individual utility maximization), whereas the agents

in the social simulation behaved adaptively. Agents made decisions by trial and error, and they learned from experiences to make better decisions. [37]

5 Discussion

In this article, different points of view into modern scientific practices that are enabled by the use of information and communication technologies have been presented. Some examples of using data and text mining methods, in particular, the self-organizing map, in studies within social and human sciences were described. In particular, the epistemological challenges related to scientific knowledge construction and communication have been addressed. One can ask what kind of future scenarios there are when scientific practices in general, and specifically in social and human science are considered.

In the area of natural sciences, *meta-analysis* has become increasingly popular. In medical research, an analysis of several studies related to a particular topic have been considered to be particularly useful (see, e.g., [38, 39]). In traditional meta-analysis, the studies are selected based on some incorporation criteria, and the variables included in the study are chosen. Meta-analysis leads to a shift of emphasis from single studies to multiple studies. It emphasizes the practical importance of the effect size instead of the statistical significance of individual studies [5]. In human and social sciences, the relevant information in scientific publications is often presented in a qualitative form describing the context and findings of the research in natural language expressions. It would be unrealistic to try to create a formalized framework for representing highly contextual and culturally grounded knowledge. On the other hand, it seems feasible to apply and develop methodologies that are used to analyze vast collections of scientific articles and model their main contents in a manner that is faithful to the complexity of the underlying phenomena. Some steps towards this direction include analysis of different conceptions (e.g. related to philosophy, see [40]), modeling of subjective conceptual spaces [14], and text mining that supports qualitative research [18].

Other applications of data and text mining include modifications to processes of reviewing research papers and project proposals. The data and text mining techniques can be combined with social simulation to create approaches for putting research teams together. For instance, it appears that the match between, for example, proposals and their evaluators can be found using pattern matching techniques applied to full text document contents.

In general, it is foreseeable that scientific practices will evolve rapidly thanks to the powerful information and communication technologies and data analysis methodologies that are available. In the social and human sciences, however, it is important that the complex and dynamical nature of the studied phenomena is properly taken into account and thus these areas require the most advanced methodological innovations in order to be successful and to avoid dangerous simplifications. One key to proper approaches may be to favor contextually rich

[5] `http://en.wikipedia.org/wiki/Meta-analysis`

representations of information. Mathematically this will mean that the methods will have to deal with very high-dimensional spaces as well as to be able to represent time-dependent and non-stationary processes that are grounded at multiple levels simultaneously.

References

[1] A. Ben Wagner. Open access citation advantage: An annotated bibliography. *Issues in Science and Technology Librarianship*, (60), 2010.

[2] D. Post and L. Votta. Computational science demands a new paradigm. *Physics Today*, 58(1):35–41, 2005.

[3] L. Vygotski. *Thought and Language*. MIT Press, 1962.

[4] Timo Honkela and Nina Janasik. Externalization and internalization of linguistic norms. In *Abstracts for Language, Culture and Mind IV (forthcoming)*, Turku, Finland, 2010. Abo Akademi University.

[5] Kimmo Koskenniemi. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis, University of Helsinki, Department of General Linguistics, 1983.

[6] J. Chandioux. MÉTÉO: un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public. *Meta*, 21:127–133, 1976.

[7] Christopher D. Manning and Heinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[8] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1), 2007.

[9] F. Sadat, G. Foster, and R. Kuhn. Système de traduction automatique statistique combinant différentes ressources. In *Conférence du traitement automatique des langues naturelles, TALN 2006*, 2006.

[10] Leigh Tesfatsion and Kenneth L. Judd, editors. *Handbook of Computational Economics, edition 1*. Elsevier, 2006.

[11] M. Cottrell, B. Girard, Y. Girard, C. Muller, and P. Rousset. Daily electrical power curves: classification and forecasting using a Kohonen map. In J. Mira and F. Sandoval, editors, *From Natural to Artificial Neural Computation. Proceedings of International Workshop on Artificial Neural Networks*, pages 1107–1113. Springer-Verlag, Berlin, Germany, 1995.

[12] Guido Deboeck and Teuvo Kohonen. *Visual Explorations in Finance with Self-Organizing Maps*. Springer-Verlag, 1998.

[13] A. Lendasse, J. Lee, E. de Bodt, V. Wertz, and M. Verleysen. *Connectionist Approaches in Economics and Management Sciences*, chapter Approximation by Radial-Basis Function networks - Application to option pricing, pages 203–214. Kluwer academic publishers, 2003.

[14] Timo Honkela, Ville Könönen, Tiina Lindh-Knuutila, and Mari-Sanna Paukkeri. Simulating processes of concept formation and communication. *Journal of Economic Methodology*, 15(3):245–259, 2008.

[15] Michael Sintek and Stefan Decker. Using TRIPLE for business agents on the Semantic Web. *Electronic Commerce Research and Applications*, 2(4):315–322, 2003.

[16] Pyry Niemelä and Timo Honkela. Analysis of parliamentary election results and socio-economic situation using self-organizing map. In *Proceedings of WSOM'09*, pages 209–218, 2009.

[17] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 2001.

[18] Nina Janasik, Timo Honkela, and Henrik Bruun. Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods*, 12(3):436–460, 2009.

[19] Marie Cottrell, Smail Ibbou, and Patrick Letrémy. Som-based algorithms for qualitative variables. *Neural Networks*, 17(8-9):1149–1167, 2004.

[20] H. Bruun, R. Langlais, and N. Janasik. Knowledge networking: A conceptual framework and typology. *VEST*, 18(3-4):73–104, 2005.

[21] H. Bruun, J. Hukkinen, K. Huutoniemi, and J Thompson Klein. *Promoting Interdisciplinary Research: The Case of the Academy of Finland*. Academy of Finland, Helsinki, 2005.

[22] Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela. A language-independent approach to keyphrase extraction and evaluation. In *Coling 2008: Companion volume: Posters*, pages 83–86, 2008.

[23] Paul Portner and Barbara H. Partee. *Formal Semantics: The Essential Readings*. Blackwell, Oxford, 2002.

[24] Richard Montague. The proper treatment of quantification in ordinary english. In J. Hintikka, J. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 221–242, Dordrecht, 1973. D. Reidel.

[25] Heinz von Foerster. *Understanding Understanding*. Springer-Verlag, New York, 2003.

[26] Richard Rorty. *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ, 1979.

[27] Terence Moore and Chris Carling. *The Limitations of Language*. Macmillan Press, Houndmills, 1988.

[28] W.V. Quine. *Word and Object*. MIT Press, 1960.

[29] E. von Glasersfeld. *Radical Constructivism. A Way of Knowing and Learning*. Falmer Press, London, 1995.

[30] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[31] Peter Gärdenfors. *Conceptual Spaces*. MIT Press, 2000.

[32] K. Hakkarainen, T. Palonen, S. Paavola, and E. Lehtinen. *Communities of networked expertise: Professional and educational perspectives*. Elsevier, Amsterdam, 2004.

[33] R. Sun. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press, 2006.

[34] Gero Schwenk and Torsten Reimer. Simple heuristics in complex networks: Models of social influence. *Journal of Artificial Societies and Social Simulation*, 11(3):4, 2008.

[35] S. Gigerenzer, P. Todd, and the ABC Research Group. *Simple heuristics that make us smart*. Oxford University Press, New York, 1999.

[36] Emmanuel Lazega. *The collegial phenomenon: the social mechanisms of co-operation among peers in a corporate law partnership*. Oxford University Press, Oxford, 2001.

[37] I. Nishizaki, H. Katagiri, and T. Oyama. Simulation analysis using multi-agent systems for social norms. *Computational Economics*, 34(1):37–65, 2009.

[38] A.J. Sutton, D.R. Jones, K.R. Abrams, T.A. Sheldon, and F. Song. *Methods for Meta-analysis in Medical Research*. John Wiley, London, 2000.

[39] C. Goutte, L. K. Hansen, M. G. Liptrot, and E. Rostrup. Feature-space clustering for fmri meta-analysis. *Human Brain Mapping*, 13:165–183, 2001.

[40] Anna-Mari Rusanen, Otto Lappi, Timo Honkela, and Mikael Nederström. Conceptual coherence in philosophy education - visualizing initial conceptions of philosophy students with self-organizing maps. In B. C. Love, K. McRae, and V. M. Sloutsky, editors, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 64–70, Austin, TX, 2008. Cognitive Science Society.