Editorial

Uwe Mönnich, Kai-Uwe Kühnberger

# Editorial

## 1 Introduction

The rise of the world-wide-web in connection with the tremendous increase of electronically available textual data of all kinds, types, genres, and forms make the scientific study of text resources a trend-setting research endeavor. The joint work of researchers trained in different disciplines and research traditions encompassing the theoretical study of properties of texts, text transformations, markup languages, query languages, and text structures, as well as the practical pursuit of archiving textual information, retrieving background knowledge from texts, and adapting dynamically ontological knowledge to new information can be considered as the birth of *text technology* as a scientific discipline.

It is not very astonishing that the triumphal elevation of hypertexts – powered by the success of the world-wide-web – as a data structure, did play an important role in making text technology a widely recognized scientific subject. Text technology as a scientific discipline has a rather short history, although its origins have their roots in classical academic research traditions like (computational) linguistics, computer science,artificial intelligence, literary sciences, and text sciences. Despite its recent emergenceas a coherent body of research text technology can easily be distinguished from theseneighboring areas and can claim to have become an autonomous discipline of its ownright. Text technology differs from classical computational linguistics and natural language processing in focusing on text as a means in itself, not on text as a container for language expressions (sentences) or text as a representation (or coding mechanism) of utterances. Moreover, text technology considers structures and layouts of texts contrary to literary history or classical linguistics and does not concentrate on finding generative principles for the question of what constitutes a text. Last but not least, it uses and develops markup standards (like XML, RDF, OWL etc.) and algorithms for statistical and symbolic computations on texts, very similarly to computer science and artificial intelligence in the area of the semantic web. But unlike computer science, text technology directs its attention on texts instead of data structures in general, therefore it attempts to structure data rather differently in comparison to, for example, the semantic web tradition, and combines ideas from structure transformations that are not at the center of interest of the computer science community.

As a research area that is crucially located between different disciplines, text technology is a strongly interdisciplinary research attempt. It combines research methodologies from the already mentioned disciplines like linguistics, computational linguistics, artificial intelligence, and computer science. Whereas the domain of interest remains inthe realm of the humanities, namely texts of all sorts and types, from a methodological point of view text technology is primarily concerned with formal sciences. With the emergence of new types of text (mainly based on the success of the world-wide-web), that are no longer linearly structured, but contain hyperlinks and multi-media data, new interaction paradigms and usability aspects are provoked and generate new needs for finding, retrieving, and editing information. Furthermore new techniques for archiving multi-modal linguistic knowledge need to be developed and implemented that have firmly established the importance of the interdisciplinary research endeavor *text technology.*

In comparison to text technology, the term *ontology* as it is used in technical disciplines has a rather different history. Picking up a term that has a history of more than two thousand years in philosophy, researchers in artificial intelligence introduced ontologies into their discipline as a means to represent conceptual background knowledge in expert systems (Brachman and Schmolze, 1985). During the further development it turned out, that many applications, most recently web applications, would strongly benefit from a sound basis on which semantic information can be coded (Daconta et al., 2003). It is a rather natural idea to integrate ontological knowledge into current text technological applications. The result is an enrichment of structural information: for example, taking annotation graphs as structural representation formalisms into account, adding ontological knowledge to annotation graphs enlarges the structural representation of text data by semantic knowledge.

The present volume is the first part of a double volume about "Ontologies in Text Technology" covering the theoretical basis of the topic. It contains a representative sample of cutting-edge work in the foundations of combining text technology and ontologies for state-of-the-art techniques of processing texts in language technology. Volume II entitled "Applications of Ontologies in Text Technology" will be published in January 2008 and will contain more applied work in the area of anaphora resolution, discourse parsing, and extracting synonymy relations and lexico-semantic classes from text.

The origins of this double volume go back to the workshop "Adaptive Ontologies on Syntactic Structures" held in conjunction with the 28th Annual Meeting of the *German Association of Linguistics* (DGfS) at the University of Bielefeld in February 2006. As a follow-up workshop, the editors organized an international workshop in Osnabrück in September 2006 entitled "Ontologies in Text Technology – Approaches to Extract Semantic Knowledge from Syntactic Information". The proceedings of this workshop contain six page papers of the participants and were published in the PICS series (Publications of the Institute of Cognitive Science). Due to the fact that with Guus Schreiber and Klaus Schulz two distinguished keynote speakers, gave inspiring talks in this workshop, the workshop attracted many internationally well-known researchers working in text and language technology. Because of the great success of this workshop the idea was conceived to provide a possibility to present the results of this workshop to a broader audience. It was decided that the participants of the workshop should be invited to submit full and extended versions of their papers for a journal publication. After a thorough further reviewing process and a revision of the accepted full articles, the result is the present double volume of the *GLDV-Journal for Computational Linguistics and Language Technology.*

## 2 The Research Unit 437 *Text Technological Information Modeling*

During the last six years the development of text technology in Germany was strongly influenced by the research unit 437 "Text Technological Information Modeling" funded by the *German Research Foundation* (DFG). This research unit is an interdisciplinary research endeavor carried by the Universities of Bielefeld, Gießen, Dortmund, Tübingen, and Osnabrück. Starting in the year 2001, this group constitutes the largest collaborative research project devoted to text technological issues and has provided the basis for text technological research in Germany. Currently this research unit is in its final funding year. In order to get a better impression of the overall project, a concise overview of the involved sub-projects of the second phase of this research unit is given:

- *Secondary Structuring of Information and Comparative Analysis of Discourse.*
  Principal Investigator: Dieter Metzing.

- *Induction of Document Grammars for the Representation of Logical Hypertextual Document Structures.*
  Principal Investigator: Alexander Mehler.

- *Text-Grammatical Foundations for the (Semi-)Automated Text-to-Hypertext Conversion.*
  Principal Investigator: Angelika Storrer.

- *Generic Document Structures in Linearly Organized Texts: Text Parsing Using Domain Ontologies and Text Structure Ontologies.*
  Principal Investigator: Henning Lobin.

- *Adaptive Ontologies on Extreme Markup Structures.*
  Principal Investigators: Uwe Mönnich, Kai-Uwe Kühnberger.

Although the research unit tries to cover all aspects of current text technological activities, it is easily possible to identify certain core aspects that play a central role in all sub-projects. Examples for such vertical topics of the whole research unit are ontologies, annotations, markup standards, and processing aspects of texts. All these topics play an important role in all participating projects. Some aspects of these vertical topics of the research unit are also represented in this double volume of the GLDV-*Journal*. The present volume focuses on the foundations of theories for developing, characterizing, coding, learning, and adapting ontological background knowledge as a crucial challenge for the semantic annotation of text documents. Some of the sub-projects of the collaborative research unit mentioned above are represented in this volume. Others will document aspects of their work in Volume II "Applications of Ontologies in Text Technology". We think that we can provide by this not only a representative documentation of text technology in general, but also a representative collection illustrating the research unit 437 in particular.

## 3 The Structure of Volume I

This first volume "Foundations of Ontologies and Text Technology" contains articles concerned with the methodological basis of using ontologies in text technology. Two aspects need to be distinguished in this context: the syntactic aspect attempts to focus on the underlying languages and data structures used for coding technologically relevant information, as for example, markup standards like XML and annotation graphs as a means to code linguistic information. Complementary to the syntactic level, the semantic aspect deals with properties of ontological knowledge for text technological applications. Both topics include aspects of learning and adaptation: learning and adaptation of ontologies is a research field that is of great importance for the future, because hand-coded ontologies are tedious, time-consuming, and expensive to create (Perez and Mancho, 2003). But also on the syntactic side, there is the need for the development of learning mechanisms: learning text types based on structural information only, without any information about their content, turns out to be possible in many cases. In the following, we will summarize major aspects of the articles included in this volume.

Lexical-semantic networks like the well-known WordNet (Fellbaum, 1998) together with its versions in other languages like RussNet or GermaNet are not only a de facto standard for several applications in text technology, but can also be seen as prototypical examples where ontological knowledge can successfully be applied in text technology. In their article "Domain Ontologies and Wordnets in OWL: Modelling Options", *Harald Lüngen* and *Angelika Storrer* question the common conversion standard to interpret synsets and lexical units of WordNet as OWL individuals. The article provides arguments for a different conceptual view, namely that synsets and lexical units need to be interpreted as concepts instead. Technically this results in a different modeling of codingWordNet ontologies in the OWL format. The authors base their claim on an evaluation of OWL representation models for WordNet variants like GermaNet combined with TermNet.

The second article of this volume "Automatic Ontology Extension: Resolving Inconsistencies" by *Ekaterina Ovchinnikova* and *Kai-Uwe Kühnberger* continues the discussion of ontologies by focusing on learning and adaptation aspects of ontologies against the background of new input. The work follows the tradition to represent ontological knowledge using description logic, therefore it considers ontology design from a logical perspective (Baader et al., 2003). Due to the fact that automatically generated and automatically updated ontologies face the problem of becoming inconsistent, the paper provides an automatic procedure for resolving occurring inconsistencies in ontology design. Potential inconsistencies in ontology design are restricted to logical ones, in particular, the overgeneralization of concepts and polysemy problems are discussed in detail. The authors propose an algorithmic solution for an automatic resolution based on the minimal non-conflicting substitute.

Related to the question of how to consistently extend ontologies by dynamic updates is the question of how the population of ontologies with existing data sources can be achieved. The article "Integration Languages for Data-Driven Approaches to Ontology

Population and Maintenance" by *Eduardo Torres Schumann*, *Uwe Mönnich*, and *Klaus Schulz* proposes a new integration language that is capable of generating new entries in large-scale ontologies based on structured data. By an intelligent user interface an efficient way to supervise the population of the ontology can be provided. The authors embed their work into a system that is able to encode large amounts of data like encyclopedic and common purpose knowledge: this large-scale knowledge base is called EFGT net (Schulz and Weigel, 2003), a precisely defined framework designed for various NLP applications.

Text technology is concerned with different types of text. Not only that different types of text have different content, often they differ significantly on a structural level as well. Concerning webpages one can distinguish, for example, homepages of scientists, blogs, or online stores from each other by structural features (Lindemann and Littig, 2006). The article "Structural Classifiers of Text Types: Towards a Novel Model of Text Representation" by *Alexander Mehler*, *Peter Geibel*, and *Olga Pustylnikov* discusses possibilities to learn text types solely on the basis of structural information without having any content information. The authors show that the document object model (DOM) can be used, in order to code structural information of texts. The authors propose different learning mechanisms for achieving this task like quantitative structure analysis (QSA) and several variants of tree kernels. The article adds also an evaluation of these learning algorithms based on a large newspaper corpus.

Graph structures play an important role in coding linguistic and textual information. Prominent examples are annotation graphs, which are used to represent multi-layered information about language, like phonological, grammatical, semantical, and pragmatical information, as well as non-linguistic information (gestures or cultural background). On the other hand, tree structures can be used in order to analyze text types. The article "Towards a Logical Description of Trees in Annotation Graphs" by *Jens Michaelis* and *Uwe Mönnich* focuses on logical descriptions of annotation graphs, one of the major data resources for text technological applications. The authors present results for characterizing a large class of annotation trees, namely, single time line, multiple tiers (STMT) models, which constitute a subclass of annotation graphs in the sense of Bird and Liberman (2001), and from which multi-rooted trees can be constructed. Besides other technical results, the article provides also a spelled-out algorithm for tree-like graph transduction from a given STMT model into a multi-rooted tree. The result is a uniform and mathematically rigorous format for the syntactic representation of annotation graphs. Taking into account that multi-rooted trees and Bird-Liberman annotation graphs play a prominent role in archiving and coding texts, this work can be considered as a theoretical basis for annotation tasks in general.

## 4 Acknowledgments

This volume would not have been possible without the help of many people. Most importantly the guest editors want to thank the editors-in-chief of the GLDV-*Journal for Computational Linguistics and Language Technology*, Alexander Mehler and Christian Wolff. Their encouragement and support in all phases of the emergence of this double volume has been irreplaceable in completing it. Furthermore we want to thank the German Research Foundation for financial support of the research unit 437 “Text Technological Information Modeling” and particularly the speaker of this research unit, Dieter Metzing.

Last but not least, the editors want to thank the program committee for their careful evaluations of the submitted papers. The quality of this volume is also a direct consequence of the work these reviewers invested. The program committee consisted of the following researchers (in alphabetical order): Irene Cramer, Thierry Declerck, Stefan Evert, Pascal Hitzler, Wolfgang Höppner, Helmar Gust, Marcus Kracht, Edda Leopold, Alessandro Moschitti, Larry Moss, Rainer Osswald, Olga Pustylnikov, Georg Rehm, Hans-Christian Schmitz, Bernhard Schröder, Uta Seewald-Heeg, Manfred Stede, Markus Stuptner, Frank Teuteberg, Yannick Versley, Johanna Völker, Armin Wegner, and Christian Wolff.

### References

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., editors (2003). *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press, New York.

Bird, S. and Liberman, M. (2001). *A Formal Framework for Linguistic Annotation.* Speech Communication, (33):23–60.

Brachman, R. and Schmolze, J. (1985). *An Overview of the KL-ONE Knowledge Representation System.* Cognitive Science, 9:171–216.

Daconta, M., Obrst, L., and Smith, K. (2003). *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management.* John Wiley and Sons.

Fellbaum, C., editor (1998). *WordNet. An Electronic Lexical Database.* MIT Press.

Lindemann, C. and Littig, L. (2006). *Coarse-Grained Classification of Web Sites by their Structural Properties.* In Proc. of WIDM’06, pages 35–42.

Perez, G. A. and Mancho, M. D. (2003). *A Survey of Ontology Learning Methods and Techniques.* OntoWeb Delieverable 1.5.

Schulz, K. and Weigel, F. (2003). *Systematics and Architectures for a Resource Representing Knowledge about Named Entities.* In Proceedings Workshop on Principles and Practice of Semantic Web Reasoning, pages 189–207.