Harald Lüngen, Alexander Mehler, Angelika Storrer

Editorial

Following the special editions 22(2) and 23(1) on *Foundations of ontologies in text-technology*, this is the third volume of the LDV-Forum to originate from activities of the research unit 437 *Text-technological modelling of information*, funded by the German Research Foundation (DFG) in its second phase 2005-2009. One of the research goals shared by four out of five subprojects within the group was the automated analysis of different types of discourse relations and the construction and evaluation of domain ontologies and lexical-semantic wordnets as knowledge sources in this task.

The subproject HyTex - *Text-Grammatical Foundations for the (Semi)-Automated Text-to-Hypertext Conversion* (Principal Investigator: Angelika Storrer) was concerned with the identification of thematic development structures in specialised texts for the purpose of hypertextualisation. The goal of the subproject *Indogram - Induction of document grammars for the Representation of Logical Hypertextual Document Structures* (Principal Investigator: Alexander Mehler) was to research methods of learning document and content structures from very large corpora of hypertext (web) documents. Finally, the subproject *SemDok - Generic Document Structures in Linearly Organised Texts* (Principal Investigator: Henning Lobin) dealt with building a text parser in the framework of Rhetorical Structure Theory for the complex text type of scientific research articles. As a joint intitiative of these three projects, a workshop entitled *Ontologies and Semantic Lexica in Automated Discourse Analysis* was held in conjunction with the *Arbeitskreis Korpuslinguistik* of the GLDV at the GLDV Frühjahrstagung in April 2007 in Tübingen.

The workshop included a most inspiring invited talk by Manfred Stede, entitled "Gewusst wie - Ontologien und Textkohärenz", which presented an insightful overview of the history of automatic text understanding systems from early knowledge-based systems within the discipline of Artificial Intelligence in the 1970s and 1980s to the rise of statistical methods in the 1990s and to more recent hybrid approaches and his own text-technological, multi-level analysis approach to text processing. A concise version of this historical overview is now contained as a part of Stede's contribution to the present volume.

Altogether, four of the five contributions to the present volume are paper versions of talks held at the workshop, namely the ones by Manfred Stede, Caroline Sporleder, Bärenfänger et al. and Diewald et al. The fifth article by Cramer et al. was additionally reviewed and included due to its immediate relevance for the field of lexical-semantic ressources in automated discourse analysis. Each paper was reviewed for the LDV-Forum by two external reviewers; as is customary, reviewers for the first round were chosen by the guest editors, and reviewers for the second round were chosen by the regular editors of the LDV-Forum.

Discourse analysis in the title of this volume, and, relatedly, *discourse structure* and *discourse relations* are used as cover terms for various types of relational structuring of text beyond the domain of sentences. Several levels of discourse structure can be identified on account of the types of relations used and the linguistic units involved in them. Let us briefly review the most important levels of discourse structure and current theoretical and practical approaches to their (automated) identification in text.

The first one is the level of coreference, or anaphora. Anaphora is a cohesive phenomenon that occurs intra-sententially as well as inter-sententially. Anaphoric relations are usually described to hold between discourse entities that are elements of the semantic model of the text world, or alternatively, between the linguistic units that are used to express them. Current anaphora resolution systems such as described in Vieira (2000) or Stuckardt (2005) use large amounts of annotated training corpora. Diewald et al.'s (pp. 74–92) contribution to this volume describes a novel, web-based multi-user annotation tool for semantic relations which can be used to produce corpora for a specialised task like anaphora resolution. In Bärenfänger et al. (pp. 49–72), an ontology of anaphoric relation types is introduced, and the interrelationship between anaphora, rhetorical relations, and thematic development is examined in a corpus study based on linguistic annotations on multiple levels.

On another level of discourse structure, which may be called lexical cohesion after Halliday and Hasan (1976), the content words occurring in a text are grouped into lexical chains based on lexical-semantic relations holding between them. Lexical chaining is a computational linguistic application first introduced by Morris and Hirst (1991). In their original algorithm, chains were derived by means of looking up lexical-semantic relations between content words in Roget's Thesaurus. More recent approaches to lexical chaining use wordnets as a knowledge source (Hirst and St-Onge, 1998), or establish semantic relatedness using terminologies or social ontologies such as Wikipedia (Mehler 2009, Waltinger et al., 2008). The third article of this volume (Cramer et al. pp. 34–47) describes experiences with the development of the lexical chainer GLexi, which derives semantic distances from GermaNet and was tested on the HyTex corpus of German specialised texts. GLexi is ultimately supposed to function as a component in the generation of topic views as an automated hypertextualisation strategy applied to a given linear text. Topic views can be regarded as an approximation to a representation of thematic development in a text.

Many researchers use the term *discourse structure* exclusively as a label for the system of coherence relations between text segments of different size, usually with propositional content. Examples of discourse relations of this type are the causal, the contrastive, or the elaboration relation. A number of discourse theories aim at describing the admissible structures of text-type-independent discourse coherence relations, notably SDRT (Asher and Lascarides, 2003), RST (Mann and Taboada, 2006), or the ULDM (Polanyi et al., 2003). Discourse coherence relations are frequently associated with lexical items called *discourse connectives*, but at the same time, many coherence relation instances in

a text lack overt cues. The contribution by Sporleder (p. 20–32) in this volume presents an evaluation of machine learning models in which lexical-semantic relations from the Princeton WordNet are used to disambiguate discourse coherence relations from SDRT that lack overt or unambiguous discourse markers. The contribution by Bärenfänger et al. (p. 49–72) explores the question how types of anaphoric relations annotated in a corpus of scientific articles can help identify instances of the RST elaboration relation.

We would like to thank both the review board for the extended abstracts submitted for the Tübingen workshop as well the reviewers of the paper versions. Without their support we would not have accomplished another edition of the LDV-Forum of such high calibre: Irene Cramer, Marcus Egg, Christiane Fellbaum, Iryna Gurevych, Anke Holler, Kai-Uwe Kühnberger, Peter Kühnlein, Lothar Lemnitzer, Henning Lobin, Vivian Raithel, Georg Rehm, Roman Schneider, Bernhard Schröder, Manfred Stede and Christian Wolff. Furthermore, we would like to thank the editors of the LDV-Forum, Alexander Mehler and Christian Wolff and their team for their support and advice. We hope that the readers of the LDV-Forum will find the included papers as interesting and illuminating as we did.

December 2008

Harald Lüngen
Alexander Mehler
Angelika Storrer

References

Literatur

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. English Language Series. Long man, London, 5 edition.

Hirst, G. and St-Onge, D. (1998). Lexical chains as representation of context for the detection and correction malapropisms. In Fellbaum, C., editor, *WordNet: An electronic lexical database.*

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Polanyi, L., van den Berg, M., and Ahn, D. (2003). Discourse structure and sentential information structure. *Journal of Logic, Language and Information*, 12:337–350.

Stuckardt, R. (2005). A machine learning approach to preference strategies for anaphor resolution. In Branco, A., McEnery, T., and Mitkov, R., editors, *Anaphora Processing: Linguistic, Cognitive, and Computational Modelling*. John Benjamins.

Taboada, Maite and Mann, William C. (2006). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Vieira, R. u. M. P. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579.

Waltinger, U., Mehler, A., and Stührenberg, M. (2008). An integrated model of lexical chaining: Applications, resources and its format. *In KONVENS 2008 - Ergänzungsband Textressourcen und lexikalisches Wissen*, pages 59–70, Berlin.

Alexander Mehler (2009): A Quantitative Graph Model of Social Ontologies by Example of Wikipedia. In: Alexander Mehler, Serge Sharoff and Marina Santini (eds.): *A Quantitative Graph Model of Social Ontologies by Example of Wikipedia. Genres on the Web: Computational Models and Empirical Studies*, Forthcoming.