

**Zentrum für interdisziplinäre Forschung (ZiF), Universität Bielefeld**

**„Processing Text-Technological Resources“**

International and Interdisciplinary Conference

**Bielefeld, 13.3.2008–15.3.2008**

## **Section 1: Text Parsing: Data Structures, Architecture, Evaluation**

Text parsing is based on theories of discourse and their principles for the analysis of relations between text segments where relations are not restricted to certain types of text. Text parsing of complex text types such as scientific journal articles requires the analysis of a document on linguistic and structural levels that go beyond traditionally employed lexical and grammatical discourse markers. The Project C1 *Generic Document Structures in Linearly Organized Texts* of the research group *Text-technological Modelling of Information* is concerned with building a text parser that takes into account discourse segments, lexical discourse markers, and generic document structures such as the logical document structure, texttype structure, and thematic structures in order to derive discourse structures on the micro- and the macro-level of a document (<http://www.texttechnology.de/SemDok/>). The target structures of the parser are discourse trees in the spirit of Rhetorical Structure Theory (RST). However, the traditional view that discourse structures are represented in an empirically adequate way as tree structures, has recently been challenged by Florian Wolf in his dissertation (MIT, Department of Brain and Cognitive Science, 2005). Consequently, the topical research questions in this field are: How should multiple analyses of text be derived and represented so that they can function as efficient resources for text parsing in a texttechnological environment? What is the best processing architecture for a text parser that processes linguistic annotations on multiple levels? Are trees or graphs adequate target data structures to represent discourse, and if they are graphs, which parts (e.g. discourse relations) could still be represented by trees and which parts exploit the full expressive power of graph structures?

Given the frequency of crossed dependencies and of nodes connected to multiple higher nodes, discourse graphs (labeled chains of graphs) have been proposed as empirically adequate data structures by Wolf & Gibson. Their approach is corpus-based and counts as the first broad evaluation of data structures for the representation of discourse coherence.

Wolf & Gibson's results were based on corpus investigations using the GraphBank annotation tool. The contribution "*Advances in discourse: Theory and annotation*" by **John Kraemer, Mark Finlayson, Denise Ichinco & Edward Gibson**, which will be presented by John Kraemer, introduces StoryBank, which is a new annotation tool that goes beyond GraphBank and is suited to investigate questions related to the role of lexical discourse markers, discourse segments, co-reference, and entities and events. New evidence from data collected with StoryBank shows that different classes of discourse relations obey distinct structural constraints.

**Manfred Stede** has researched multi-level annotations of analyses of argumentative texts in various projects at the University of Potsdam, where the so-called Potsdam Commentary Corpus has been created, completed with an infrastructure of annotation and retrieval tools. His contribution *addresses "Multi-Level Representations: Implications for Text Parsing."*

In the discussion part, results from the C1 project will be presented by **Henning Lobin and Harald Lungen** in a talk entitled "*Processing texttechnological resources in discourse parsing*".

## **Section 2: Recognition of Thematic Structures: Methods, Resources, and Applications**

The automated recognition of thematic structures is an important issue for many applications of text technology and information extraction. The *HyTex* project (B1), which coordinates this section, is concerned with this issue in the context of text-to-hypertext conversion (cf. <http://www.hytex.info>). In this context, the project investigates methods to automatically generate topic chains and topic views on hypertext bases. Since the approach developed in this project is based on lexical chaining, the focus of the workshop will be on the following topics:

(1) Methods and resources to calculate semantic relatedness/similarity; particularly measures used for lexical and thematic chaining.

(2) Evaluation of lexical resources (particularly the Princeton WordNet and GermaNet) used for lexical and thematic chaining.

The following presentations will address selected aspects of these topics and frame the subsequent discussion in this section.

**Irene Cramer, Marc Finthammer & Angelika Storrer:** "*Generating topic chains and topic views: Experiments using GermaNet*".

This talk presents GLexi, a lexical chainer implemented for German corpora, which uses GermaNet as well as Google co-occurrence counts as a resource for the calculation of semantic relatedness/similarity. It will be discussed, how GLexi may be used in the context of text-to-hypertext conversion to generate topic chains and topic views. In addition, studies are reported that evaluated the performance of the topic chains and views with respect to manually annotated data.

**Christiane Fellbaum:** "*Towards a better lexical resource for NLP*". The Princeton WordNet is an important resource for many NLP applications. Christiane Fellbaum discusses results on psycholinguistic experiments aimed at increasing WordNet's connectivity by identifying syntagmatic links among synsets in ways that do not introduce biases and limitations inherent to traditional, systematic, introspectively defined relations. These results raise interesting questions as to the nature of semantic relations, semantic similarity, and human conceptual organization.

**Graeme Hirst, Saif Mohammad & Meghana Marathe:** "*Semantic distance measures with distributional profiles of coarsened-grained concepts*".

Graeme Hirst presents a hybrid measure of semantic distance based on distributional profiles of concepts inferred from text corpora. Because the measure

is based on naturally occurring text, it is able to find word pairs that stand in non-classical relationships not found in WordNet. He will report the results of experiments using this measure in text segmentation and he will discuss some recent challenges to the utility of lexical chains.

**Anke Holler:** *"A psycholinguistic perspective on the relevance of discourse structure to anaphora resolution"*.

The interpretation of discourse anaphora plays a central role in natural language processing applications. Anke Holler presents a psycholinguistic study that investigates inasmuch discourse structural information affects the way intersentential anaphora are resolved. The questionnaire study aims at an empirical verification of the so-called Right Frontier Constraint first proposed by Polanyi (1988).

### **Section 3: From Textual Data to Ontologies, from Ontologies to Textual Data**

The goal of this section is to distinguish main steps for the extraction of semantic relations from texts, as well as to develop solutions for the adaptation of already generated ontologies, in the case new information is available. Ontology development and applicability are research topics of all projects of the research unit *Text-technological Modelling of Information*, as documented in the Workshop 12 "Adaptive Ontologies on Syntactic Structures" held in conjunction with the 28<sup>th</sup> Annual Meeting of the "Deutsche Gesellschaft für Sprachwissenschaft" in Bielefeld (February 2006) and the International Workshop "Ontologies in Text-Technology" that took place in Osnabrück in September 2006 (<http://www.cogsci.uniosnabrueck.de/~ott06>). Both workshops were organized by the principal investigators of the C2 project *Adaptive Ontologies on Extreme Markup Structures* (cf. <http://www.text-technology.de>).

Due to the fact that the manual development of major ontologies on the basis of textual data is very expensive, semi-automatic methods are therefore used. Very often, however, the results are not free from contradictions, i.e. ontologies become logically inconsistent causing problems for validation, for ontology alignment, and for standard inference processes. Furthermore the problem of losing information during inference processes can occur. This section aims to cover the sketched field by a talk of **Kai-Uwe Kühnberger, Jens Michaelis, Uwe Mönnich and Tonio Wandmacher** entitled "*Adaptation of Ontological Knowledge from Structured Textual Data*", by a talk of **Steffen Staab** entitled "*On understanding the collaborative construction of conceptualizations*", and by a talk of **Alessandro Oltramari** entitled "*On the boundary between Ontologies and Lexical Resources*". Additionally, a short position paper by **Philipp Cimiano** will be presented in the discussion section entitled "*Ontology Learning Revisited: what have we been doing?*".

## Section 4: Multidimensional Representations: Solutions for Complex Markup

Digitalized texts and their processing based on external resources are the objects of research of the project *Secondary Information Structuring and Comparative Analysis of Discourse* (cf. <http://www.texttechnology.de/Sekimo/>). Text-external resources are implemented grammars, lexica, ontologies, amongst others. Their application generates layers of annotations, which are the input for markup unification, information integration and analysis. Anaphora resolution is taken as a test case for the application of external resources and for the processing of markup. These aspects are focussed upon in the contribution of **Daniela Goecke and Maik Stührenberg**, "*Integrated linguistic annotation models and their application in the domain of antecedent detection*".

Research on architectures for multidimensional markup is carried out in different projects and domains. Wouter Alink will present his research in the domain of Digital Forensics. Given his experiences with *XIRAF* (**I**nformation **R**etrieval **A**pproach to digital **F**orensics), a program environment facilitating the integration of external resources for multidimensional markup generation, as well as with the query language *XQuesta*, a special language for search and information integration of multidimensional markup, Wouter Alink will compare approaches in domains of texttechnology and digital forensics. Objects of analysis are data and processes of a computer, and information integration is supposed to produce high quality evidence of illegal computational actions. There are common methods, e.g. the use of stand-off annotation as a solution for multidimensional markup, similar methods, e.g. automatic feature extraction as a kind of document enrichment and annotation, and differences in the treatment of digital documents; and there is a new discipline in search for scientific standards: **Wouter Alink**, "*A comparison of markup problems found in multilevel text-processing and digital forensics*".

The use of XML as a standard for data representation and interchange has been highly successful. However, complex markup generates problems the solution of which is still an active research field (cf. the contributions to the recent *Extreme Markup* workshops). "Recent work on markup has emphasized the importance of overlapping structures, but little progress has been made towards validation of such structures." (Sperberg-McQueen, 2006) Two approaches have been proposed: level by level validation of complex markup ("tree-by-tree validation") and level integrating validation ("interaction constraints of trees"). These approaches will be a topic of this section: **Michael Sperberg-McQueen**, "*Containment and dominance in Goddag structures*"; **Oliver Schonefeld**, "*Validation of overlapping markup*".

## Section 5: Document-Structure Learning: Procedures for Multiply Structured Data and Web-Based Document-Types

Approaches to text mining face an enormous challenge given the dynamics of structure formation in the WWW. One reason is that patterns of web-based documents – which are in a rapid flux compared to *text* types – emerge spontaneously and rapidly during the short history of the WWW. This typological dynamics is accompanied by the temporal dynamics of the life cycle of web documents which includes frequent modifications, inconsistent markup and implicit delimitations from the surrounding web. Thus, any approach to web *content* mining relies on approaches to web *structure* mining expressive enough to master this structural dynamics. This integrative view of content and structure mining is at the core of the Project A4 *Induction of Web Genre Document Grammars* (cf. <http://www.texttechnology.de/Indogram/>) of the research group *Text Technological Modelling of Information*. It focuses on textual as well as on hypertextual data structures ranging from sequences of lexical and sentential units via hierarchical logical document structures to the point of graph-like document networks induced by hyperlinks. In this sense, multiply structured data are at the core of Section 5 as initiated by Project A4. The central text technological research question posed in this context is: *Which machine learning architecture is best suited to map which type of (e.g. tree- or graph-like) document structure?* The section “Document-Structure Learning” presents approaches to learning and retrieving such structured data. This includes the following contributions: **Gerhard Paaß** focuses on *Logical Document Structures* (LDS) by example of textual units. He tackles a central object of structure learning, that is, poly-hierarchies of the kind of the LDS. In this context, Paaß provides a machine learning method based on conditional random fields by which columns, tables and related LDS segments are segmented. This is a central prerequisite for web structure mining as the majority of web data is implicitly structured thereby suffering from the so-called tag-abuse problem.

The combination of XML information retrieval and machine learning is a central research topic of **Ludovic Denoyer**. In his contribution he focuses on mapping semi-structured documents – that is, documents in which structure is already annotated in part – on XML schemata as a prerequisite of retrieval from document corpora. Denoyer provides a ML-method for learning document structures based on reinforcement learning and, thus, bridges the areas of structure modeling and information retrieval. While Gerhard Paaß and Ludovic Denoyer focus on structures of textual and hypertextual units, respectively, the exploration of *content*-related units is dealt with by **Gerhard Heyer**. He contributes an approach to *content* mining from large corpora of natural language texts whose vocabularies are mapped onto complex networks as an alternative to the predominant architecture of Euclidean semantic spaces. In this sense, Heyer provides means to enhance purely structure-oriented approaches to document modeling.

The integration of content and structure modeling is at the core of the contribution of **Alexander Mehler and Armin Wegner**. They present a unified approach to content and structure mining by example of web documents which tackles a central problem of learning from web-based data.

This relates to the delimitation of instances of web-based document types sub-

ject to their graph-inducing link structure. Accordingly, Mehler & Wegner present a machine learning method for semi-supervised hypertext zoning.

The typological question about the range of web document types and their relation to text technological modeling is raised by **Marina Santini**. She starts from a functional notion of web genre in order to classify central tasks in modeling, segmenting and relating web-based document types. In this sense, Santini provides a typological background of the research presented in Section 5.

As Section 5 is directed towards an integration of document structure and content learning it intends to connect yet unrelated areas in text-technology. This research is highly relevant for the development of document models, which map the dynamics of web-based communication.

## **Section 6: Sustainability of Text-Technological Resources**

Long-term usability of important resources, however, is a serious problem, therefore sustainability of resources are becoming a standard task in language documentation and language description. In this section sustainability of language resources will be in focus as well as sustainability considerations in the field of text-technological research. The state of the art and perspectives for a scientific community sharing web-based standards will be presented by **Gary Simons**: "*Ensuring the sustainability of language resources*".

Solutions to sustainability problems in linguistics are the central goal of the project *Sustainability and Linguistic Data* of the German Research Council (DFG) connected to more than 40 projects of three linguistic longterm research units ("Sonderforschungsbereiche", SFB). There are two main research directions explored, the development of common representation formats for heterogeneous data, compatible with an ontology for linguistic terms and concepts, on the one hand and best practice rules for acquisition, access and archiving of linguistic resources. Research of the linguistic sustainability project will be presented by **Andreas Witt**: "*Sustainability of language resources: A project report*".

Solutions for sustainability problems are relevant for the Research Unit *Text-technological Modelling of Information* in at least three respects. They are directly relevant for resources produced and shared by the five research projects. Furthermore, methods, approaches and experiences of these projects may contribute to current sustainability research, e.g. as to the use of complex markup or as to the combination of markup with ontologies or as to formal and complexity considerations of annotation systems. However, it is not yet clear to which extent sustainability requirements for linguistic resources and for text-technological resources are identical, since part of the latter are not only texts and their annotations but also hypertexts and websites ("text nets"), aspects to be discussed in this section.