Sustainability of Text-Technological Resources

Maik Stührenberg[1], Michael Beißwenger[2], Kai-Uwe Kühnberger[3], Harald Lüngen[4]
Alexander Mehler[1], Dieter Metzing[1], Uwe Mönnich[5]

[1]Universität Bielefeld, [2]Technische Universität Dortmund, [3]Universität Osnabrück,
[4]Justus-Liebig-Universität Gießen, [5]Universität Tübingen

Abstract

We consider that there are obvious relationships between research on sustainability of language and linguistic resources on the one hand and work undertaken in the Research Unit "Text-Technological Modelling of Information" on the other. Currently the main focus in sustainability research is concerned with archiving methods of textual resources, i.e. methods for sustainability of primary and secondary data; these aspects are addressed in our work as well. However, we believe that there are additional certain aspects of sustainability on which new light is shed on by procedures, algorithms and dynamic processes undertaken in our Research Unit.

The Research Unit "Text-Technological Modelling of Information"

The Research Unit 437 "Text-Technological Modelling of Information" is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) and consists of five projects.[1] The funding started in 2002, since October 2005 the Group is in its second period lasting until late 2008. Before we will go into detail about the observations we made about relationships between research on sustainability of language and linguistic resources and the work carried out in our Research Unit, we will first introduce the projects that take part in it.

A2: Sekimo The topic of the project Sekimo is the integration of heterogenous linguistic resources which can be divided into annotated textual documents on the one hand and grammars, parsers, lexicons or ontologies – amongst others – on the other hand. The project focusses on the application and integration of the latter mentioned on raw texts or pre-annotated documents. For accomplishing this task two architectures have been developed: a Prolog fact based approach, developed in the first period of the project and described in detail in Witt (2004), and an XML-based approach which makes use of the SEKIMO GENERIC FORMAT (SGF) in conjunction with a native XML database system (Stührenberg and Goecke, 2008). Both architectures can be used to examine relationships between modelling units derived from different annotation layers without the need for markup unification – which could leed to overlapping problems. The exemplifying application of these architectures focusses the analyses of anaphoric relations which are of high relevance for projects in the inner context of the research group (e.g. when measuring the value of anaphoric relations as cues for rhetorical relations) and for external cooperations. The corpus under investigation is based on the corpus of German scientific articles of the C1 (SemDok) project and was extended with German newspaper texts. All texts are annotated with multiple annotation layers, including a document structure layer (developed by the projects B1 and C1), a morphological and syntactic layer (provided by the commercial tagger software MACHINESE SYNTAX from Connexor Oy, which is used in other projects of the Research Group as well), a discourse entity layer (automatically generated), and the cohesive layer containing the semantic relations. For the latter task the web-based annotation tool SERENGETI was developed (Stührenberg et al., 2007), which allows for both high quality and quantity annotation of texts carried out by a large number of users. In addition the comparison of annotations on the same text made by different users is possible as well. Ongoing work on SERENGETI includes the support of the SFG and the possibility of user-defined annotation schemas.

A4: Indogram The topic of the project Indogram is the automatic induction of probabilistic document grammars as models of hypertext types or web genres, respectively. It develops an algorithm for learning the internal structure of web documents as instances of web genres (e.g. conference websites, personal academic home pages or weblogs). The project starts from the idea that an appropriate web genre model gets its validity to the degree to which it clarifies the relation of *explicit* (*visible*) or *manifesting website* structure and *implicit* (*hidden*) or *manifested web genre* structure. Thus, beyond tagging genre labels to websites or pages the learning of genre-related web document structures is a major project goal. A central observation which makes this kind of structure mining a challenge beyond classical approaches to text mining is that hypertext graphs of the same type are distributed according to a multidimensional power-law (Mehler et al., 2007b). Thus, there are no typical page-based web genre structures so that classical approaches to structure learning cannot be applied. In order to solve this task various structure-related classifiers were developed which operate on the level of textual (Mehler et al., 2007a) and hypertextual (Mehler, 2008) structures. They show that classifiers of structure come into reach with a very low space and a moderate time complexity. Further,

---

[1]More information can be observed at `http://www.text-technology.de/`

A4 has developed a two stage model of hypertext types which combines an SVM tagger of genre constituents with a HMM of their networking. For this task, A4 explores *structural* features (on the level of the logical document structure), *lexical* features (including named entities) and HTML features. Thus, A4 has built a classifier which utilizes heterogeneous linguistic resources. Further, in order to master the dynamics and semantic diversity of websites, A4 has developed an approach to topic labeling based on social ontologies and, thus, combines content and structure mining. This model has been utilized for implementing a prototype for hypertext zoning, that is, an algorithm which delimits websites no longer in terms of physical (URL-related) features, but in terms of their content and function.

B1: HyTex On the basis of a corpus of documents from two scientific domains (hypertext and text technology)[2], project B1 develops and evaluates innovative strategies for handling conceptual problems of the so-called text-to-hypertext conversion. The approach is coherence-based (Kuhlen, 1991) and aims at generating hypertext views which provide the selective reader instant access to all textual units that he or she may need for a proper understanding of the current hypertext node and, thus, make selective reading and browsing more efficient and more convenient than would be possible with print media (Lenz and Storrer, 2006; Storrer, 2008). The strategies developed in the project process markup information from different annotation layers. XML document grammars have been developed for:

- the document structure layer (applying an annotation scheme derived from DOCBOOK, which was developed in cooperation with project C1; cf. Lenz and Lüngen (2004));

- the terms and definitions layer, on which occurrences as well as definitions of technical terms in the documents are annotated (Storrer and Wellinghoff, 2006; Lenz et al., 2006; Wellinghoff, 2006);

- the thematic structure layer (applying an annotation scheme based on the typology of thematic progression according to Zifonun et al. (1997) §C6; cf. Lenz and Storrer (2003));

- the cohesion layer, on which various types of text-grammatical information are annotated (e.g. co-reference, connectives, text-deictic expressions; Holler (2003a; Holler (2003b; Holler et al. (2004)).

Additional linguistic information was provided by morphosyntactic annotations automatically assigned by the KAROPARS technology (Müller, 2004).
Besides coherence-based strategies for segmentation and linking on the document level, the HyTex approach also comprises strategies for providing hypertext users

with navigation devices that support the reconstruction of domain-specific knowledge as well as thematical orientation while browsing the hypertext version of a scientific document:

- On the one hand and with special respect to the needs of users which are "semi-experts" in a certain domain, the project built up TERMNET, a WORDNET-style semantic net which describes the technical terminology specific to the respective domains (Beißwenger, 2008). On the presentation level, TERMNET is used for generating glossary views that are linked to the term occurrences in the corpus.

- On the other hand, the project develops topic-based linking strategies which use a GERMANET-based lexical chaining approach as a resource (cf. (Cramer and Finthammer, 2008)) and which aim at generating topic views – thematic indices based on text-grammatical information constructed of a selection of topic items – as an additional navigation and orientation device.

C1: SemDok The goal of the SemDok project is to develop a text parser (discourse parser) for a complex text type in the framework of Rhetorical Structure Theory (RST, Mann and Thompson (1988); Marcu (2000)). The linguistic features for discourse interpretation of scientific research articles are firstly derived from a discourse marker lexicon with about 100 entries encoding features (e.g. induced relation, directionality, discourse segment level, frequency, and others) of lexical discourse markers (i.e. conjunctions and discourse adverbs). Secondly, a development corpus of German research journal articles was compiled and subsequently annotated on several levels of linguistic analysis corresponding to the output of pre-processing components for logical document structure analysis (Lenz and Lüngen, 2004), text type structure analysis (Bärenfänger et al., 2006), initial discourse segmentation (Lüngen et al., 2006a), and lexical discourse marker annotation. Each level was added as a separate XML annotation layer in the framework of XML-based multi-layer annotation (Witt, 2004), with document grammars formulated as XML schemas. Several articles were also annotated according to RST-HP, which is the XML application that serves as the target structure of the SemDok parser (Lüngen et al., 2006b). It utilises the XML document tree to represent an RST discourse tree. The SemDok hierarchy of rhetorical relations called "RRSet" is an adaptation of previously suggested relation taxonomies to the analysis of scientific research articles. It is formalised in the web ontology language OWL, as an extension of the work described in (Goecke et al., 2005), cf. Bärenfänger et al. (2007). The perl program RS3TOHP converts manual annotations of RST structure built with the RSTTool by O'Donnell (2000) into the RST-HP format. Additionally, morphological and syntactic annotations are provided using the already mentioned commercial tagger MACHINESE SYNTAX. The development corpus and its annotations will

be made available as soon as the relevant legal issues are clarified with the publisher of the research articles.

C2: Ontologies Text technologically based information modelling is confronted with two main problems for which there are no solutions in formal linguistics. On the one hand, there is the phenomenon of markup-structures which are despite of their character similar to trees not presentable in classical techniques of tree grammars. This lack of presentability is due to the possibility of unbounded branchings and the appearance of secondary relations. On the other hand, the dynamic aspect of web-oriented ontologies is a challenge which can not be refuted by the means of methods of dynamic logics and their linguistic incarnation. The C2 project attempts to extract ontological knowledge from syntactically given information (coded in annotation graphs), in order to expand ontologically coded information. On the syntactic side, a major goal of the project is the logical and complexity theoretic characterization of certain types of annotation graphs, such as the well-known Bird-Liberman graphs (Bird and Liberman, 2001). In Michaelis and Mönnich (2007), 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, and from which multi-rooted trees can be constructed. On the semantic side, it is a matter of fact that automatic as well as semi-automatic procedures for the expansion of ontologies (triggered by information coded in annotation graphs) contain errors of different types. These errors can range from structural and user-defined inconsistencies to logical contradictions (Haase and Stojanovic, 2005). In a series of papers, members of the C2 project provide algorithmic solutions for resolving automatically certain types of logical inconsistencies in ontology design relative to various types of description logics (cf. (Ovchinnikova and Kühnberger, 2006a; Ovchinnikova et al., 2007; Ovchinnikova and Kühnberger, 2007).

1. Overview and Preliminary Distinctions

We consider that there are obvious relationships between research on sustainability of language and linguistic resources on the one hand and work undertaken in the Research Unit "Text-Technological Modelling of Information" on the other. Furthermore, we see new relationships that merit to be explored in more detail. An important aspect of sustainability research is the long-term availability of resources, either for basis research or for applications. Sustainability research has many facets, the following aspects may be distinguished: aspects related to primary data, secondary data and category systems (cf. Section 2.); aspects related to procedures for these data and category systems (cf. Section 3.); aspects related to process properties in a long-term perspective cf. Section 4.); and aspects related to a community of experts and non-experts agreeing to work with shared standards (cf. Section 5.). Work undertaken in the Research Unit is concerned

with these aspects with the exception of aspects related to process properties, which presupposes an organizational framework in order to guarantee sustainability, for example the constant actualization of the process organization and the continuous adjustment to changed goals and basic conditions.

2. Data and Category Systems
2.1. Sustainability of Primary and Secondary Data

The building and usage of corpora are important aspects of text-technological research. Examples of corpora that were built in the Research Unit and which are partially available online were provided by all projects. For assuring sustainability of primary and secondary data, the usage of XML (in favour of proprietary formats) can be considered as key issue. XML-based modelling of information has been the base line of the Research Unit. This is reflected both by the explicit usage of XML in the A2 (Sekimo) project and the usage of XML representations for coding document and discourse structure and metadata in other projects. The explicit usage of XML in the Sekimo project consists of several format descriptions for annotation schemas (both in XML DTDs and XML Schema Descriptions, XSD) and results in the development and implementation of the generic representation format SGF (SEKIMO GENERIC FORMAT) as basis for a generic architecture (in conjunction with a database backend, either native XML or relational, cf. Stührenberg et al. (2006)). The XSD-based SGF consists of a base layer that uses a standoff approach (Thompson and McKelvie, 1997) for storing multiple annotated data. An arbitrary number of annotation layers, separated via distinct namespaces, can be imported into the base layer (Stührenberg and Goecke, 2008). In contrast to similar approaches such as the pivot format of the LAF (Linguistic Annotation Framework, cf. Ide et al. (2003)) or PAULA (Potsdamer Austauschformat für linguistische Annotation, cf. Dipper (2005)), SGF uses only a single file to store multiple annotations on a single or even multiple files and imports all sorts of annotation schemas (including graph based). The format is designed to stick as accurately as possible to the imported annotation layer – including the possibility of validating its content – and uses standard XQuery (instead of introducing extensions, cf. Alink et al. (2006)) for analyzing relationships between elements derived from different layers.[3]

Further, XML-based formats for poly-hierarchical hypertext structures as well as document networks (Mehler and Gleim, 2006) were developed in the Indogram project, including text internal structures down to the level of dependency trees (Pustylnikov and Mehler, 2008; Pustylnikov et al., 2008). A basic requirement of structure-oriented annotations is the flexibility and adaptability of the annotation format in use. This requirement has been met by further implementing a graph theory-related format in conjunction with a text-

---

[3]Possible relations are described in Witt et al. (2005).

technological database operating thereon in the Indo-gram project (Gleim et al., 2007a), and – in the Sekimo project – by the before mentioned generic representation format SGF and its employment in conjunction with native XML databases.

Project C2 chose to represent automatically extracted lexical-semantic patterns in a standard description logic format (which can be considered as a syntactic variant of OWL).

Standards for metadata are considered to be a necessity for text technological applications, this applies to the research carried out in our Research Unit, too. Recommendations regarding the use of metadata standards (Dublin Core, cf. DCMI Usage Board (2006) OLAC, cf. Simons and Bird (2003) and IMDI, cf. IMDI (ISLE Metadata Initiative) (2003) the latter one for multimodal corpora) were given in an examination of several annotation standards for structuring and representing textual corpora (DocBook, TEI, CES and XCES) carried out in the Sekimo project (Stührenberg, 2007). OLAC metadata is used both in the projects Sekimo (exclusively, imported into the SGF) and SemDok (in addition to newly developed metadata sets). Options guaranteeing an open access to our corpora (e.g. Open Access[4] or Creative Commons[5]) are still under discussion, however, as stated in the description of the projects, some corpora are available to the public at present.

2.2. Sustainability of Category Systems

Category systems (like ontologies) can be used as a mediator between different sets of annotation (Schmidt et al., 2006). However, one has to assure that the category system is sustainable as well. One way to guarantee sustainability of category systems is the introduction of standardized formal ontologies specifying the categories and, hopefully, mediating between different category systems. The overall (and long-term) goal of this mediation is interoperability between different ontological resources. Among the standards that are used to ensure sustainability of linguistic and text-technological resources, ontologies and the Web Ontology Language OWL (W3C Web Ontology Working Group, 2004) play an important role. OWL is a W3C recommendation, and as an XML application, it offers a standardised formalism. As an ontology language it allows for a formal description of the semantics of XML tag sets. There are two ways in which ontologies are employed to ensure aspects of sustainability:

1. Reference ontologies as instruments to ensure the sustainability of linguistic tag sets.
   Linguistic or text-technological resources (corpus annotations) are made interoperable by mapping the category sets employed in them (annotation schemes) onto a formal ontology that has been introduced as a proposal for standardisation in the domain. An example of such an ontology is GOLD

(Farrar and Langendoen, 2003). Moreover, the reference ontology developed in the SFB 441[6] is designed to link domain ontologies that represent the syntax and morphology annotation schemes of three different research projects (Chiarcos, 2007).

2. Construction of ontological resources.
   Linguistic resources other than annotations, especially lexical-semantic resources are represented (pro-actively, or by retroactive conversion) in a standardised ontology formalism, such as the OWL versions of the Princeton WordNet (van Assem et al., 2006).

With respect to these two directions, several domain-specific ontologies have been developed in the research unit: Regarding the first point, existing ontological standards for linguistic domains such as GOLD currently provide universal concepts for morphological and syntactic categories. Those categories however, for which resources were constructed in the research unit, are mostly found on the textual levels of linguistic analysis, i.e. above syntax and morphology, e.g. logical text objects, discourse entities, discourse relations, co-reference, lexical chains, topic chains, and text type structure categories. Presently, no ontological standardisations of discourse categories are available, but within the research unit, a proposal for a an ontology of discourse units and relations (rhetorical relations and anaphoric relations) as an extension of the GOLD approach has been put forward (Goecke et al., 2005).

In order to research ways of making general-language and domain-specific wordnets interoperable, different aspects of modelling wordnets in OWL have been researched in a cooperation of the C1 (SemDok) and the B1 (HyTex) project. In doing so, several resources have been constructed in OWL: Firstly, the terminological wordnet TermNet, in which terminology from domains of hypertext research and text technology is represented on the basis of an wordnet model that has been extended for terminologies (Beißwenger, 2008; Selzam, 2008). Secondly, the GermaNet resource; thirdly, an integrated version of TermNet and a subset of GermaNet that have been connected in OWL via so-called plug-in relations (Lüngen and Storrer, 2007; Kunze et al., 2007; Lüngen et al., 2008).[7] Other domain-specific ontologies that have been developed in the Research Unit are a framework for integrating lexical ontologies as a resource of semantic annotation of documents (Goecke et al., 2007b; Mehler et al., 2007c) and a lexical-semantic ontology based on automatically extracted patterns from heterogeneous resources, where a special focus concerns the integration of primary data into one homogeneous database of hypotheses (Krumnack et al., 2007). A remaining challenge is the development of procedures for adapting and merging different ontologies (cf. Section 3.).

2.3. Availability of Methods and Tools

Often a sustainable use of methods and tools is prevented by the fact that documentation or source code is not made available to the public. In case of our Research Unit, documentations of markup specifications developed for diverse layers of linguistic annotation are already available online: e.g. in terms of thematic structure (Lenz and Storrer, 2003), coreference phenomena (Holler, 2003a; Holler, 2003b; Holler et al., 2004), term definitions in text (Storrer and Wellinghoff, 2006; Lenz et al., 2006; Wellinghoff, 2006), and an annotation schema for annotating anaphoric relations (Goecke et al., 2007a). The sources of the web-based annotation tool SERENGETI[8] (Stührenberg et al., 2007) will be made publicly available online under the GPL (GNU Public License) before the end of the Research Unit together with the corresponding documentation. Further, the ARIADNE SYSTEM for the development, maintenance and statistical analysis of large-scale multimodal corpora[9] (Gleim et al., 2007a), the SCIENTIFIC DESKTOP for the semantic analysis of web documents[10] (Waltinger et al., 2008) and the WEBCEP SYSTEM[11] (Gleim et al., 2007b) for the development and maintenance of web genre corpora are available online.

3. Procedural Aspects

Procedural aspects of sustainability are based on the idea that long-term archiving of text-technological resources cannot be reduced to a static saving of data.

3.1. Learning and Induction of Ontological Systems

The hand-coded development of ontologies, relevant for the interoperability of category systems, is a tedious, time-consuming, and expensive task. Therefore automatic procedures for the extraction of knowledge and the learning of ontologies play a central role now and in the future. Inductive methods from machine learning for explorative data analysis and the build-up of ontologies and text technological resources were developed (Mehler et al., 2007c; Waltinger et al., 2008). Architectures for the integration of heterogenous linguistic resources and partial solutions for the automatic extraction of heterogeneous data sources and the transformation of these data in a format that allows uniform querying were developed as well (Krumnack et al., 2007; Goecke et al., 2007b).

3.2. Dynamics and Consistency Preservation of Ontological Resources

Due to the fact that sustainability of textual data requires the possibility of extensions of ontological resources, consistency problems of such resources can be considered as a central problem. Several types of problems can be distinguished and were partially solved: overgeneralizations of concepts in case that non-monotonic extensions of the underlying data basis are necessary (Ovchinnikova and Kühnberger, 2006a), undergeneralizations of concepts (Ovchinnikova and Kühnberger, 2006b), and polysemy problems (Ovchinnikova and Kühnberger, 2007). These results were tested on example ontologies (Ovchinnikova et al., 2007). Following the state-of-the-art to code ontological knowledge in description logics results in the task to develop different resolution algorithms relative to the chosen description logic. Due to the fact that different description logics have different expressive strengths and different constructors can be used to form new concepts, there is no easy way to expand a known algorithm for a certain DL to another more expressive DL. Resolution strategies for certain types of inconsistencies can be provided for mild extensions of the attributive language family.

3.3. Formal Properties of Data Structures

Only a formally precise characterization of the underlying data structures of repositories ensures that algorithmic solutions can be found for sustainability questions and the comparability of different data formats. Purely structural properties of trees and graphs can be learnt in the framework of quantitative structure analysis and graph kernel methods which are successfully used to classify document types (Dehmer and Mehler, 2007; Mehler et al., 2007a). A logical characterization of annotation graphs as well as a constructive procedure in order to algorithmically transform the annotation graphs developed by Bird & Liberman (Bird and Liberman, 2001) into multi-rooted trees was developed by (Michaelis and Mönnich, 2007). In Mönnich and Kühnberger (2007), this approach is embedded into a broader context of text technological research. Connected with the formal specification of the underlying coding formats with respect to their complexity theoretic properties are import-export interfaces for standards and representation formalisms like RDF, OWL (in its different versions), SKOS etc.

4. Process Perspective

In a process perspective of sustainability the interrelationships between different actors or institutions may be focussed upon. This may be illustrated by four actors and their relationships introduced by Gary Simons in his talk at the conference "Processing Text-Technological Resources" at Bielefeld (Germany) in March this year.[12] First, there is the *creator* who brings a resource into existence, preferably according to aspects introduced in Section 2.. Second, there is the *curator* or *archiver* who takes on the responsibility to sustain the necessary conditions for use, preferably according to aspects introduced in Section 3., especially interoperability. Third, there is the *aggregator* who

---

http://www.wordnets-in-owl.de.

[8]http://coli.lili.uni-bielefeld.de/serengeti/

[9]http://varda.coli.uni-bielefeld.de:8080/Ariadne/

[10]http://www.scientific-workplace.org

[11]http://ariadne.coli.uni-bielefeld.de:8080/WikiCEP/

---

[12]The short abstract is available at http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/

takes care of the web-accessibility of resources archived at different places, preferably according to advanced search procedures, i.e. according to aspects introduced in Section 3.. Forth, there is the *user* expecting that resources of interest to him are and will be discoverable; or there is a *community* of users, introduced in Section 5., that may influence the process of sustainability at different stages.

5.   Community of Experts and Non-Experts

A central aspect of sustainability is the existence of a community and a complex network of valuable cooperation, i.e. a community agreeing to work with shared standards as well as with procedures of interoperability, accepting the web-based collaboration with experts and non-experts and exploiting the web as a field of global cooperation (Simons, 2007). An example in this sense is the cooperation of the Sekimo project with the international projects "Anaphoric Bank"[13] and "AnaWiki" (Poesio and Kruschwitz, 2008) contributing to its corpora, representation format and the architecture of the before mentioned web-based annotation tool SERENGETI. A special wiki that supports scientific collaboration with respect to the exchange and maintenance of treebanks[14] was implemented in the Indogram project (Pustylnikov and Mehler, 2008), and by means of a scientific desktop[15] parts of the procedural output of our research were made available.

6.   Conclusion

As stated by Simons (2007) the concept of sustainability has many facets. The work undertaken in the Research Unit "Text-Technological Modelling of Information" account for most of them and proposes additional aspects of sustainability regarding procedures, algorithms and dynamic processes. In addition, the cooperation between different projects in the Research Unit, the use of shared corpora, methods and tools has prevented multiple implementation and reduced the overall amount of work in these fields. Apart from the projects that take part in the Research Unit the publicly available access to most of the documentation and tools can help other interested projects in the same way and can contribute to a community building process.

7.   References

Wouter Alink, Raoul Bhoedjang, Arjen P. de Vries, and Peter A. Boncz:. 2006. Efficient XQuery Support for Stand-Off Annotation. In *Proceedings of the 3rd International Workshop on XQuery Implementation, Experience and Perspectives, in cooperation with ACM SIGMOD*, Chicago, USA, Juni.

Maja Bärenfänger, Mirco Hilbert, Henning Lobin, Harald Lüngen, and Csilla Puskàs. 2006. Cues and constraints for the relational discourse analysis of complex text types – the role of logical and generic document structure. In Candy Sidner, John Harpur, Anton Benz, and Peter Kühnlein, editors, *Proceedings of the Workshop on Constraints in Discourse*, pages 27–34. National University of Ireland, Maynooth, Ireland.

Maja Bärenfänger, Henning Lobin, Harald Lüngen, and Mirco Hilbert. 2007. Using OWL ontologies in discourse parsing. In Kai-Uwe Kühnberger and Uwe Mönnich, editors, *OTT'06 - Ontologies in Text technology: Approaches to Extract Semantic Knowledge from Structured Information. Series Publications of the institute of Cognitive Science (PICS) 1*, Osnabrück.

Michael Beißwenger. 2008. TERMNET — ein terminologisches Wortnetz im Stile des Princeton Wordnet. Technical report of the B1 Project. `http://www.hytex.info/`.

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1–2):23–60.

Christian Chiarcos. 2007. An ontology of linguistic annotation: Word classes and morphology. In *Proceedings of DIALOG 2007, Bekasovo/Moscow*.

Irene Cramer and Marc Finthammer. 2008. An evaluation procedure for word net based lexical chaining: Methods and issues. In *Proceedings of the Global WordNet Conference 2008, Szeged, Hungary*. `http://www.inf.u-szeged.hu/projectdirs/gwc2008/`.

DCMI Usage Board. 2006. DCMI Metadata Terms. DCMI Recommendation, Dublin Core Metadata Initiative.

Matthias Dehmer and Alexander Mehler. 2007. A new method of measuring the similarity for a special class of directed graphs. *Tatra Mountains Mathematical Publications*, 36:39–59.

Stefanie Dipper. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin, Deutschland.

Scott Farrar and D. Terence Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *GLOT International*, 7(3):97–100.

Rüdiger Gleim, Alexander Mehler, and Hans-Jürgen Eikmeyer. 2007a. Representing and maintaining large corpora. In *Proceedings of the Corpus Linguistics 2007 Conference, Birmingham (UK)*.

Rüdiger Gleim, Alexander Mehler, Matthias Dehmer, and Olga Pustylnikov. 2007b. Aisles through the category forest – utilising the wikipedia category system for corpus building in machine learning. In Joaquim Filipe, José Cordeiro, Bruno Encarnação, and Vitor Pedrosa, editors, *3rd International Conference on Web Information Systems and Technologies (WEBIST '07), March 3-6, 2007, Barcelona*, pages 142–149, Barcelona.

Daniela Goecke, Harald Lüngen, Felix Sasaki, Andreas Witt, and Scott Farrar. 2005. GOLD and Discourse: Domain- and Community-Specific Extensions. In *Proceedings of the E-MELD Workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Language Resources*, Cambridge, Massachusetts.

Daniela Goecke, Anke Holler, and Maik Stührenberg. 2007a. Koreferenz, Kospezifikation und Bridging: Annotationsschema. Technical report of the A2 Project.

Daniela Goecke, Maik Stührenberg, and Tonio Wandmacher. 2007b. Extraction and representation of semantic relations for resolving definite descriptions. extended ab-

---

PTTR/abstracts/Abstract-Simons.pdf

[13] `http://www.anaphoricbank.org`

[14] `http://ariadne.coli.uni-bielefeld.de/wikis/treebankwiki/index.php5/Main_Page`

[15] `http://www.scientific-workplace.org`

stract. In Uwe Mönnich and Kai-Uwe Kühnberger, editors, *OTT'06. Ontologies in Text Technology: Approaches to Extract Semantic Knowledge from Structured Information*, volume 1-2007 of *Publications of the Institute of Cognitive Science (PICS)*, pages 27–32. Institute of Cognitive Science, Osnabrück, January.

P. Haase and L. Stojanovic. 2005. Consistent evolution of OWL ontologies. In *Proceedings of the Second European Semantic Web Conference*, pages 182–197, Lissabon.

Anke Holler, Jan-Frederik Maas, and Angelika Storrer. 2004. Exploiting coreference annotations for text-to-hypertext conversion. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 651–654, Lissabon.

Anke Holler. 2003a. Koreferenz in Hypertexten: Anforderungen an die Annotation. *Osnabrücker Beiträge zur Sprachwissenschaft*, 68:9–29.

Anke Holler. 2003b. Spezifikation für ein Annotationsschema für Koreferenzphänomene im Hinblick auf Hypertextualisierungsstrategien. Technical report of the B1 Project. `http://www.hytex.info/`.

Nancy Ide, Laurent Romary, and Eric de la Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*, Edmunton.

IMDI (ISLE Metadata Initiative). 2003. Metadata Elements for Session Descriptions. version 3.0.4. Reference Document, MPI, Nijmegen, October.

Ulf Krumnack, Ekaterina Ovchinnikova, and Tonio Wandmacher. 2007. LexO - constructing a lexical ontology from heterogenous resources. In *Proceedings of the OntoLex'07 Workshop at ISWC*, Busan, Korea.

Rainer Kuhlen. 1991. *Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Springer, Berlin.

Claudia Kunze, Lothar Lemnitzer, Harald Lüngen, and Angelika Storrer. 2007. Repräsentation und Verknüpfung allgemeinsprachlicher und terminologischer Wortnetze in OWL. *Zeitschrift für Sprachwissenschaft*, 26:267–290.

Eva Anna Lenz and Harald Lüngen. 2004. Dokumentation: Annotationsschicht: Logische Dokumentstruktur. Technical report of the B1 Project. `http://www.hytex.info/`.

Eva Anna Lenz and Angelika Storrer. 2003. Annotationsschicht: Thematische Strukturen/Themenentwicklung. Technical report of the B1 Project. `http://www.hytex.info/`.

Eva Anna Lenz and Angelika Storrer. 2006. Generating hypertext views to support selective reading. In *Digital Humanities 2006. Conference Abstracts. Paris-Sorbonne, 5–9 Juli 2006*, pages 320–323. `http://www.hytex.info/050_publikationen/LenzStorrer87.pdf`.

Eva Anna Lenz, Michael Beißwenger, and Sandra Wellinghoff. 2006. Annotationsschicht: Definitionen und Termverwendungsinstanzen. Technical report of the B1 Project. `http://www.hytex.info/`.

Harald Lüngen and Angelika Storrer. 2007. Domain ontologies and wordnets in OWL: Modelling options. *LDV Forum*, 22(2):1–19.

Harald Lüngen, Henning Lobin, Maja Bärenfänger, Mirco Hilbert, and Csilla Puskàs. 2006a. Text parsing of a complex genre. In *Proceedings of the Conference on Electronic Publishing (ELPUB)*, pages 247–256, Bansko, Bulgaria.

Harald Lüngen, Csilla Puskàs, Maja Bärenfänger, Mirco Hilbert, and Henning Lobin. 2006b. Discourse segmentation of German written text. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*, pages 245–256, Åbo, Finland. Springer.

Harald Lüngen, Claudia Kunze, Lothar Lemnitzer, and Angelika Storrer. 2008. Towards an integrated OWL model for domain-specific and general language wordnets. In *Proceedings of the 4th Global Wordnet Conference (GWC 2008)*, pages 281–296.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text*, 8(3):243–281.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summariszation*. MIT Press, Cambridge MA.

Alexander Mehler and Rüdiger Gleim. 2006. The net for the graphs – towards webgenre representation for corpus linguistic studies. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 191–224. Gedit, Bologna.

Alexander Mehler, Peter Geibel, and Olga Pustylnikov. 2007a. Structural classifiers of text types: Towards a novel model of text representation. *LDV Forum*, 22(2):51–66.

Alexander Mehler, Rüdiger Gleim, and Armin Wegner. 2007b. Structural uncertainty of hypertext types. An empirical study. In *Proceedings of the Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP", September, 30, 2007, in conjunction with RANLP 2007, Borovets, Bulgaria*, pages 13–19.

Alexander Mehler, Ulli Waltinger, and Armin Wegner. 2007c. A formal text representation model based on lexical chaining. In Peter Geibel and B. J. Jain, editors, *Proceedings of the KI 2007 Workshop on Learning from Non-Vectorial Data (LNVD 2007) September 10, Osnabrück*, pages 17–26, Osnabrück. Universität Osnabrück.

Alexander Mehler. 2008. Structural similarities of complex networks: A computational model by example of wiki graphs. *Appears in: Applied Artificial Intelligence*.

Jens Michaelis and Uwe Mönnich. 2007. Towards a logical description of trees in annotation graphs. *LDV Forum*, 22(2):68–83.

Uwe Mönnich and Kai-Uwe Kühnberger, editors. 2007. *OTT'06. Ontologies in Text Technology: Approaches to Extract Semantic Knowledge from Structured Information*, volume 1-2007, Osnabrück, January. Institute of Cognitive Science.

Frank Henrik Müller. 2004. Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z). Technical report. `http://www.sfs.uni-tuebingen.de/tupp/dz/stylebook.pdf`.

Michael O'Donnell. 2000. RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253 – 256, Mitzpe Ramon, Israel.

Ekaterina Ovchinnikova and Kai-Uwe Kühnberger. 2006a. Adaptive ALE-TBox for extending terminological knowledge. In A. Sattar and B. H. Kang, editors, *AI 2006. Proceedings of the 19th ACS Australian Joint Conference on Artificial Intelligence (LNAI 4304)*, Lecture Notes in Artificial Intelligence, pages 1111–1115. Springer.

Ekaterina Ovchinnikova and Kai-Uwe Kühnberger. 2006b. Aspects of automatic ontology extension: Adapting and regeneralizing dynamic updates. In M. Orgun, editor, *Advances in Ontologies. Proceedings of the Australasian Ontology Workshop (AOW 2006), Conferences in Research*

and Practice in Information Technology, volume 72, pages 52–60.

Ekaterina Ovchinnikova and Kai-Uwe Kühnberger. 2007. Automatic ontology extension: Resolving inconsistencies. *LDV Forum*, 22(2):19–33.

Ekaterina Ovchinnikova, Tonio Wandmacher, and Kai-Uwe Kühnberger. 2007. Solving terminological inconsistency problems in ontology design. *International Journal of Interoperability in Business Information Systems (IBIS)*, (4):65–80.

Massimo Poesio and Udo Kruschwitz. 2008. Anawiki: Creating anaphorically annotated resources through web cooperation. Submitted to LREC 2008.

Olga Pustylnikov and Alexander Mehler. 2008. Towards a uniform representation of treebanks: Providing interoperability for dependency tree data. In *Proceedings of First International Conference on Global Interoperability for Language Resources (ICGL 2008), Hong Kong SAR, January 9-11*.

Olga Pustylnikov, Alexander Mehler, and Rüdiger Gleim. 2008. A unified database of dependency treebanks. integrating, quantifying & evaluating dependency data. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech (Morocco).

Thomas Schmidt, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. 2006. Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In *Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*, Lansing, Michigan.

Bianca Selzam. 2008. Modellierung des TERMNET in OWL. Technical report of the B1 Project. `http://www.hytex.info/`.

Gary Simons and Steven Bird, 2003. *OLAC Metadata*. OLAC: Open Language Archives Community.

Gary Simons. 2007. Doing linguistics in the 21st century: Interoperation and the quest for the global riches of knowledge. In *Proceedings of the "Toward the Interoperability of Language Resources" Workshop, LSA Summer Institute*, Stanford University, July.

Angelika Storrer and Sandra Wellinghoff. 2006. Automated detection and annotation of term definitions in german text corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Angelika Storrer. 2008. Mark-up driven strategies for text-to-hypertext conversion. In *Linguistic Modelling of Information and Markup Languages*. Spinger, Dordrecht.

Maik Stührenberg and Daniela Goecke. 2008. Integrated linguistic annotation models and their application in the domain of antecedent detection. *Submitted to Balisage 2008*.

Maik Stührenberg, Andreas Witt, Daniela Goecke, Dieter Metzing, and Oliver Schonefeld. 2006. Multidimensional markup and heterogeneous linguistic resources. In David Ahn, Erik Tjong Kim Sang, and Graham Wilcock, editors, *Proceedings of the 5th EACL Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, pages 85–88, Trento. EACL.

Maik Stührenberg, Daniela Goecke, Nils Diewald, Irene Cramer, and Alexander Mehler. 2007. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 140–147, Prag, Juni. Association for Computational Linguistics. `http://acl.ldc.upenn.edu/W/W07/W07-1523.pdf`.

Maik Stührenberg. 2007. Texttechnological standards – an overview. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007*, pages 157–166, Tübingen. Gunter Narr Verlag.

Henry S. Thompson and David McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97: The next decade – Pushing the Envelope*, pages 227–229, Barcelona.

Marc van Assem, Aldo Gangemi, and Guus Schreiber. 2006. Conversion of WORDNET to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*.

W3C Web Ontology Working Group. 2004. OWL Web Ontology Language. Set of seven specifications. W3C Recommendation, World Wide Web Consortium.

Ulli Waltinger, Alexander Mehler, and Gerhard Heyer. 2008. Towards automatic content tagging: Enhanced web services in digital libraries using lexical chaining. In *4rd International Conference on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal*. Barcelona.

Sandra Wellinghoff. 2006. Manuelle Annotation definitorischer Textsegmente incl. Guidelines Phase i und ii. Technical report of the B1 Project. `http://www.hytex.info/`.

Andreas Witt, Daniela Goecke, Felix Sasaki, and Harald Lüngen. 2005. Unification of XML Documents with Concurrent Markup. *Literary and Lingustic Computing*, 20(1):103–116.

Andreas Witt. 2004. Multiple hierarchies: New Aspects of an Old Solution. In *Proceedings of Extreme Markup Languages*.

Gisela Zifonun, Ludger Hoffmann, and Bruno Strecker. 1997. *Grammatik der deutschen Sprache*. deGruyter, Berlin/New York.