Published in: Mehler, Alexander/Kühnberger, Kai-Uwe/Lobin, Henning/Lüngen, Harald/ Storrer, Angelika/Witt, Andreas (eds.): Modeling, Learning, and Processing of Text-Technological Data Structures. - Berlin/Heidelberg: Springer, 2011. pp. 1-11. (Studies in Computational Intelligence 370)

# Chapter 1
# Introduction: Modeling, Learning and Processing of Text-Technological Data Structures

Alexander Mehler, Kai-Uwe Kühnberger, Henning Lobin, Harald Lüngen, Angelika Storrer, and Andreas Witt

## 1.1 Textual Units as Data Structures

Researchers in many disciplines, sometimes working in close cooperation, have been concerned with modeling textual data in order to account for texts as the prime information unit of written communication. The list of disciplines includes computer science and linguistics as well as more specialized disciplines like computational linguistics and text technology. What many of these efforts have in common

Alexander Mehler
Computer Science and Mathematics, Goethe-Universität Frankfurt, Senckenberganlage 31, D-60325 Frankfurt am Main, Germany
e-mail: Mehler@em.uni-frankfurt.de

Kai-Uwe Kühnberger
Institute of Cognitive Science, Universität Osnabrück, Albrechtstraße 28, D-49076 Osnabrück, Germany
e-mail: kkuehnbe@uos.de

Henning Lobin
Applied and Computational Linguistics, Justus-Liebig-Universität Gießen, Otto-Behaghel-Straße 10D, D-35394 Gießen, Germany
e-mail: Henning.Lobin@germanistik.uni-giessen.de

Harald Lüngen
Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, R5, 6-13, D-68161 Mannheim, Germany
e-mail: luengen@ids-mannheim.de

Angelika Storrer
Institute for German Language and Literature, Technische Universität Dortmund, Emil-Figge-Straße 50, D-44221 Dortmund, Germany
e-mail: angelika.storrer@tu-dortmund.de

Andreas Witt
Institut für Deutsche Sprache, Zentrale Forschung, R 5, 6-13, D-68161 Mannheim, Germany
e-mail: witt@ids-mannheim.de

is the aim to model textual data by means of abstract data types or data structures[1] that support at least the semi-automatic processing of texts in any area of written communication.[2]

Generally speaking, an abstract data type is a mathematical model of a certain range of data together with operations defined on that model in such a way that they can be performed automatically on the data [2]. From this point of view, natural language texts are a very special sort of data that requires very specific data structures for being processed automatically – of course, there is no single data structure that can model all aspects of natural language texts. A central characteristic of this task is structural uncertainty.

## Structural Uncertainty

In order to understand this notion, take the example of a tree-like model of text structure as proposed, for example, by *Rhetorical Structure Theory* (RST) [7]. Any text can be made an object of operations (e.g., linkage of text spans) related to this data structure by virtue of interpreting (e.g., delimiting) its constituents (e.g., text spans). However, due to the semiotic nature of natural language texts, this interpretation is, in principle, open, that is, not necessarily determined by the data itself. Moreover, the relevant context that allows for determining this interpretation is not necessarily known in advance, nor fixed once and for all. Take the example of rhetorical structure as modeled in RST: on the one hand, there is disagreement on the range of rhetorical relations that can actually hold between text spans [8]. On the other hand, there is uncertainty about which relation actually holds between a given pair of spans – even if the set of rhetorical relations is fixed in advance – a problem known as inter-annotator disagreement [3]: often, humans diverge in their interpretation of the same data with respect to the same data structure.

In other words, the way textual data is structured is not necessarily clear from the data itself. It may be the result of semantic or even pragmatic interpretations that are ultimately carried out by humans going beyond (and possibly far beyond) the linguistic context of the data to be interpreted. As a consequence, the structure of a given text as an instance of a given data structure can be very *uncertain* as this structure does not need to be reflected by the text's constituents in an obvious way.[3]

Thus, the semi-automatic or even manual annotation of textual data, that is, its informational enrichment to make its structure explicit according to the underlying data structure, is a central task of text technology and related disciplines. This also includes linking textual data with other linguistic resources. Any such annotation and linkage, whether done manually, semi-automatically, or fully automatically, would support various tasks of text processing (e.g., information extraction [10], text categorization [9], text mining [4], text summarization [6], topic modeling

[1] In this chapter, we use these terms interchangeably.

[2] Throughout this volume, we concentrate on written communication.

[3] This sort of structural uncertainty should not be confused with the notion of semi-structured data [1, 10].

[5], or discourse parsing [8]). Any of these tasks requires expressive data structures in conjunction with efficient operations that together allow for moving closer to the goal of *automating* text processing.

This book, "*Modeling, Learning and Processing of Text-Technological Data Structures*", deals with such data structures. Here we focus on theoretical foundations of representing natural language texts as well as on concrete operations of automatic text processing. Following this integrated approach, the present volume includes contributions to a wide range of topics in the context of processing of textual data. This relates to the learning of ontologies from natural language texts, annotation and automatic parsing of texts as well as the detection and tracking of topics in texts and hypertexts. In a nutshell, the book brings together a wide range of approaches to procedural aspects of text technology as an emerging scientific discipline. It includes contributions to the following areas:

- formalizing annotations of textual units
- extracting knowledge and mining ontologies from texts
- building lexical and terminological resources
- machine learning of document structures
- classifying and categorizing texts
- detecting and tracking topics in texts
- parsing discourse

This book addresses researchers who want to get familiar with theoretical developments, computational models and their empirical evaluation in these fields of research. It is intended for all those who are interested in standards of representing textual data structures, the use of these data structures in various fields of application (such as topic tracking, ontology learning and document classification) and their formal-mathematical modeling. In this sense, the volume concerns readers from many disciplines such as text and language technology, natural language processing, computational linguistics and computer science.

## 1.2 Overview of the Book

### *1.2.1 Text Parsing: Data Structures, Architecture and Evaluation*

Part I of the volume focuses on the automatic analysis of text structure. By analogy to the analysis of sentence structure, it is often called text or discourse parsing. Fundamental aspects of text parsing are the data structures used, principles of system architecture, and the evaluation of these parsing systems.

The first chapter on "*The MOTS Workbench*" by Manfred Stede and Heike Bieler deals with the standardization of processing frameworks for text documents – an important issue for language technology for quite some time. The authors examine one particular framework, the MOTS workbench, and describe the overall architecture, the analysis modules that have been integrated into the workbench, and the user interface. After five years of experience with this workbench, they provide a

critical evaluation of its underlying design decisions and draw conclusions for future development.

The second chapter, "*Processing Text-Technological Resources in Discourse Parsing*" by Henning Lobin, Harald Lüngen, Mirco Hilbert and Maja Bärenfänger, investigates discourse parsing of complex text types such as scientific research articles. Discourse parsing is seen from a text-technological point of view as the addition of a new layer of structural annotation for input documents already marked up on several linguistic annotation levels. The GAP parser described in this chapter generates discourse structures according to a relational model of text structure, Rhetorical Structure Theory. The authors also provide an evaluation of the parser by comparing it with reference annotations and with recently developed systems with a similar task. In general, both chapters show that text or discourse parsing is no longer of purely experimental interest, but can yield useful results in the analysis of huge amounts of textual data. This will eventually lead to parsing applications that pave the way to new generations of content-based processing of documents in text technology.

### *1.2.2 Measuring Semantic Distance: Methods, Resources, and Applications*

The determination of semantic distance between lexical units is crucial for various applications of natural language processing; in the context of text technology semantic distance measures were used to reconstruct (and annotate) cohesive and thematic text structures by means of so-called lexical chains. The two contributions of Part II deal with methods and resources to determine semantic distance and semantic similarity in different application contexts.

In their chapter, "*Semantic distance measures with distributional profiles of coarse-grained concepts*", Graeme Hirst and Saif Mohammad first provide an overview of NLP applications using such measures. Then, they group the various approaches to calculate semantic distance into two classes: (1) resource-based measures which determine semantic distance by means of the structure of lexical resources such as thesauruses or word-nets; (2) distributional measures which compare distributional profiles of lexical units generated on the basis of text corpora. The authors list the strengths and limitations of the two measure classes and propose, as an alternative, a hybrid method which calculates distributional profiles not for word forms but for coarse-grained concepts defined on the basis of Roget-style thesaurus categories, disambiguating words attached to more than one concept with a bootstrapping approach. The evaluation results discussed in Section 3 of this chapter indicate that their concept-based hybrid method (using the BNC as a corpus and the Macquarie Thesaurus as a lexical resource) performs considerably better than the word-based distributional approach. However, the performance is still not at the level of the best resource-based measure obtained by using the Princeton WordNet as the lexical resource. However, not all languages dispose of resources with the coverage and quality of the Princeton WordNet. The authors show that for such

languages, a good alternative might be an extension of their method which links the concepts of the English thesaurus to a bilingual lexicon with English as the target language. This can then generate concept-based distributional profiles for the lexicon's source language. This extended method was tested for German, using the bilingual lexicon BEOLINGUS and the taz-Corpus as resources. In the comparative evaluation, presented in Section 5.2 of the chapter, the extended method performed even better than the resource-based approach using the German word-net-style resource GermaNet. In their final section, the authors show how another extension of the concept-based method may help to determine different degrees of antonymy between pairs of lexical units in text corpora.

The Princeton WordNet has proven to be a widely-used and valuable lexical resource not only for computing semantic distance but for a broad range of other natural language processing applications. However, some approaches profit from complementing the part-of-speech-specific WordNet relations by cross-part-of-speech links between semantically similar and strongly associated concepts (like [dog] and [to bark], or [sky] and [blue]). In their chapter "*Collecting Similarity Ratings to Connect Concepts in Assistive Communication Tools*", Sonya Nikolova, Jordan Boyd-Graber, and Christiane Fellbaum describe such an application context: the structuring of the vocabulary in assistive communication tools, such that people with aphasia can retrieve words that express the concepts they want to communicate. In the multi-modal visual vocabulary component for these tools, concepts are represented in combination with pictures and sounds. With each concept being mapped to a core set of the Princeton WordNet, navigating through the vocabulary can be improved using WordNet's semantic and lexical relations between words and concepts. The authors identify the need to establish additional links between concepts which are strongly associated with each other in a specific context, based on human judgments on the strength of association between disambiguated words. Since experiments to gain such judgments in controlled studies with trained persons have proven to be time-consuming and expensive, the authors have developed and tested an alternative method using Amazon Mechanical Turk. The results of their experiments indicate that this method is indeed feasible for gathering a large number of association ratings at low cost and in a short amount of time, provided that reliability checks are applied to filter out invalid ratings.

### *1.2.3 From Textual Data to Ontologies, from Ontologies to Textual Data*

Part III, "*From Textual Data to Ontologies, from Ontologies to Textual Data*", contains chapters that focus on semantic issues in text technology. Centered on the current standard of using ontologies for the coding of conceptual knowledge, this part covers semantic resources and cognitive constraints in the process of ontology creation by presenting a formal specification of the semantics of current markup standards and by proposing a general framework for the extraction and adaptation of ontological knowledge in text-technological applications.

Alessandro Oltramari's chapter "*An Introduction to Hybrid Semantics: the Role of Cognition in Semantic Resources*" argues, in a detailed way, for the consideration of cognitive structures and cognitive constraints in semantic technologies and, in particular, the process of ontology creation. The author argues that semantic approaches in text technology need to be enriched by modules that are informed by the cognitive structure of conceptualizations.

The chapter "*Modal Logic Foundations of Markup Structures in Annotation Systems*" by Marcus Kracht shows that it is useful to study connections between Markup languages and logical characterizations of these languages. The chapter shows, for example, that it is possible to derive complexity results of the query language XPath by simply transferring well-known model theoretic results of *Propositional Dynamic Logic* (PDL) to XPath.

The third chapter entitled "*Adaptation of Ontological Knowledge from Structured Textual Data*" by Tonio Wandmacher, Ekaterina Ovchinnikova, Uwe Mönnich, Jens Michaelis, and Kai-Uwe Kühnberger presents a general framework for the extraction of semantic knowledge from syntactically given information. The authors describe the transformation of this information to a logical representation and the adaptation of ontological knowledge using this new information. The framework builds on many well-known technologies and tools as, for example, WordNet and FrameNet. Further, it also builds on reasoning in description logic.

### *1.2.4 Multidimensional Representations: Solutions for Complex Markup*

Text enrichment by the substantial addition of markup is one of the characteristics of the application of text-technological methods. Part IV – "*Multidimensional representations: Solutions for complex markup*" – addresses problems related to the creation, interpretation, and interchange of richly annotated textual resources. This part includes one chapter devoted to general problems of annotated documents, and two case studies that reveal insights into specific aspects of creating and representing annotations. The chapter "*Ten problems in the interpretation of XML documents*" by C. M. Sperberg-McQueen and Claus Huitfeldt examines questions related to the semantics of document markup. The problems addressed in this chapter are related to a formalization of the interpretation of annotations. The aim of this formalization is, as it is quite often in computational linguistics and text technology, to enhance the specification of algorithms for an automatic processing of potentially richly annotated resources. The methodology presented is based on a mapping from XML annotations to predicates of first-order logic. As the title indicates, the main part of the chapter focuses on the problems which arise when transforming markup into formulas of a logical calculus. The problems discussed concern the arity of statements, the form of inference rules, deictic reference to other parts of the document, the inheritance of properties and how to override them, the treatment of a union of properties with conflicting values, the treatment of milestone elements, the definition of the universe of discourse, the occurrence of definite descriptions and multiple

references to the same individual, and the recording of uncertainty and responsibility. For each of these items, a description of the problem is presented along with a proposal of a corresponding solution.

The second chapter in Part IV is entitled "*Markup Infrastructure for the Anaphoric Bank: Supporting Web Collaboration*". The authors Massimo Poesio, Nils Diewald, Maik Stührenberg, Jon Chamberlain, Daniel Jettka, Daniela Goecke, and Udo Kruschwitz focus on the creation of specific high-quality annotations, and their publication. A collaborative approach is taken for the annotation process described in this chapter. Two different methods have been developed to support annotators. The first method is based on the special purpose editor Serengeti, a web application for modern browsers. Tools like this facilitate the process of annotation considerably, but they require expert knowledge. Since non-expert users could also provide high-quality annotations, a second method was also used. The annotations of the non-experts are produced by means of a game in which, e.g., a user tries to find the correct antecedents of an anaphora. All the annotated corpora are accessible through the web. The structure of the annotated resources is based on an XML-conformant markup scheme that makes use of the stand-off technique. This format is imported into a relational database system that allows for fast access to the data.

Part IV closes with the chapter "*Integrated Linguistic Annotation Models and their Application in the Domain of Antecedent Detection*" by Andreas Witt, Maik Stührenberg, Daniela Goecke, and Dieter Metzing. It deals with potential benefits of using information about the logical document structure for the task of anaphora resolution. Even though the investigations of these effects show only a weak influence of the logical document structure for the task (as reported in this chapter), the findings give insights into complex markup and multidimensional representations. The ways to integrate different information types in text resources are discussed at length. The discussion is based on a corpus study with a focus on logical document structure and anaphoric relations in texts. To do this study, the texts in the corpus were annotated according to the annotation schemes for these two different levels of information. The level of logical document structure was annotated according to a document grammar that uses elements from the wide-spread annotation schemes XHTML and DocBook. The other level covers the annotation of anaphoric elements, antecedents and the types of anaphoric relations. The semi-manual annotation of this anaphora level was carried out with the help of the special purpose editor Serengeti. From the point of view of multidimensional representation by means of complex markup, this chapter on the one hand presents techniques that allow for the integration of heterogeneous types of annotations in a single representation. On the other hand it presents a corpus study that investigates the interaction between diverse levels of information. The methodology described could also be adapted to examine the existence of interrelations between different linguistic levels.

### 1.2.5 Document Structure Learning

The chapters in Part V of the volume deal with document structure learning. One of the chapters focuses on "classical" texts, all other chapters deal with web documents. In this way, Part V includes models of learning the structure of texts and of hypertexts. In both cases, the *Logical Document Structure* (LDS) of a textual unit is used as a reference point for learning. In the case of texts, the LDS can be identified with their hierarchical division into sections, subsections, etc. down to the level of sentences and their lexical constituents. In the case of hypertexts, things are more complicated since hyperlinks give rise to non-hierarchical, network-like structures.

Note that the LDS is used as the reference point of many text-linguistic models that focus, for example, on rhetorical, argumentative or thematic structures. However, approaches to information retrieval that aim to go beyond the bag-of-words approach are likewise in need of models of logical document structure that can be easily induced from text instances. The chapter "*Machine Learning for Document Structure Recognition*" by Gerhard Paaß and Iuliu Konya describes approaches in this field of research. Based on the notion of a *Minimum Spanning Tree* (MST), the chapter describes an algorithm of layout-based document analysis that processes images of document pages to identify their LDS. As a matter of fact, this approach to *logical layout analysis* is highly important in the field of digitizing historical documents. However, a central problem of structure recognition relates to the variability of layout-based cues of document structuring. In order to tackle this challenge, approaches to machine learning are needed that deal with the uncertainty of layout-related manifestations of the LDS. Following this idea, the chapter reviews and describes approaches to document structure recognition that utilize *Conditional Random Fields* (CRF) as a learning method. Both classes of approaches, the MST- and the CRF-based approaches, are discussed in terms of their $F$-measure-related evaluation.

The variety of layout structures is one source of the uncertainty about the structure of non-digitized documents. A related problem concerns the variety of formats that are used to represent already digitized documents, say, by means of (X)HTML, XML-DTD, or XML Schema. Moreover, for a wide range of semi-structured documents on the web, which have multiple sources and undergo frequent changes, one has no access to the underlying document schema (if such a schema exists at all). The chapter "*Corpus-Based Structure Mapping of XML Document Corpora: A Reinforcement Learning based Model*" by Francis Maes, Ludovic Denoyer, and Patrick Gallinari addresses this kind of structural variety. Starting from a document-centric perspective, they introduce an algorithm for automatically mapping documents that vary only by their format onto a mediating schema, which expresses the structural unity of these input documents. The algorithm for aligning documents by their structure works on a set of pairs of input-output documents and, thus, is supervised. The chapter provides an extensive evaluation of this approach by means of five different corpora including a corpus of documents from Wikipedia. These experiments show that generic models of learning web document structure are possible. This, in

turn, focuses on one of the challenges of exploring information from the web that is restricted by the variety of document formats and schemata in use.

What makes the web unique in terms of document structure is its hyperlink-based structuring. That is, as instances of webgenres (i.e., types of web documents by analogy to text types), websites usually consist of several pages that are connected by hyperlinks. From a formal point of view, such documents can be seen as a special class of graphs with an internal hierarchical structure that is superimposed by graph-inducing links. Starting from this notion, the chapter "*Learning Methods for Graph Models of Document Structure*" by Peter Geibel, Alexander Mehler and Kai-Uwe Kühnberger describes two approaches to learning web document structures. First, it describes a range of kernel-based approaches that utilize structure-related kernels to build supervised classifiers of webgenres. The chapter then adopts quantitative structure analysis in order to arrive at an unsupervised classifier of the same range of webgenres. Using a corpus of three webgenres, the chapter provides empirical evidence into the learnability of hyperlink-based document structures on the level of websites.

A central aspect of learning web document structures is given by their internal and external structuring. More specifically, instances of webgenres are manifested by page-level units as well as by units across the border of single pages. Consequently, when trying to automatically delimit instances of webgenres, one has to process document-internal as well as document-external structures, whether hyperlink-based or not. The chapter "*Integrating Content and Structure Learning: A Model of Hypertext Zoning and Sounding*" by Alexander Mehler and Ulli Waltinger tackles this task. By integrating web content and structure mining, it introduces a classifier of page-internal, webgenre-specific staging together with an unsupervised algorithm of content tagging that utilizes Wikipedia as a social-semantic resource. In this way, the chapter focuses on the task of hypertext zoning, that is, of delimiting webgenre instances based on their content and structure. The chapter ends by outlining an approach to estimate bounds of thematic sounding in Wikipedia.

### *1.2.6 Interfacing Textual Data, Ontological Resources and Document Parsing*

The three chapters in Part VI deal with the extraction of semantic relations from text, the evaluation of measures of the semantic relatedness of lexemes using linguistic resources, and the modeling of WordNet-like resources in a formalism for the representation of ontologies. Semantic relations include lexical-semantic relations like antonymy and meronymy, but also a more general semantic relatedness relation sometimes called association, or evocation.

The chapter "*Exploring Resources for Lexical Chaining: A Comparison of Automated Semantic Relatedness Measures and Human Judgments*" by Irene Cramer, Tonio Wandmacher, and Ulli Waltinger gives an overview of 16 different measures of semantic relatedness using four types of linguistic resources in their calculations. They categorize the measures according to three types, that is, net-based measures,

distributional measures, and Wikipedia-based measures, and evaluate them by comparing their performance with human judgments on two lists of word pairs. The results show that among the three types of measures, distributional measures perform better than the other two types, while in general the correlations of all 16 measures with human judgments are not high enough to accurately model lexical cohesion as perceived by humans. In their conclusion, Cramer et al. argue that more research is needed in order to come up with a sounder theoretical foundation of semantic relatedness and to determine what kind of resource should be employed for what kind of task. To further these directions of research, they argue for the definition of a shared task by the research community.

The chapter "*Learning Semantic Relations from Text*" by Gerhard Heyer describes semantic relations in terms of the traditional structuralist notions of syntagmatic and paradigmatic relations between words. Accordingly, wordforms stand in a paradigmatic relation if their global contexts (statistically relevant syntagmatic relations captured in a co-occurrence set) are similar according to some (distributional) similarity measure. Semantic relations such as the hyponymy relation can then be derived by applying linguistic filters or constraints on the similarity measure. Two applications of iterative filtering and refining global contexts of words are exemplified, namely word sense disambiguation, and the identification of (near) synonymy. Heyer also sketches a language-independent, modular, web-service-oriented architecture for interactively learning semantic relations.

One type of resource frequently used in the calculation of semantic relatedness is lexical-semantic networks such as the Princeton WordNet for English. Recently, suggestions have been made to represent wordnets in formalisms designed for the representation of ontologies in order to provide better interoperability among lexical-semantic resources and to make them available for the Semantic Web. In their chapter – "*Modelling and Processing Wordnets in OWL*" – Harald Lüngen, Michael Beißwenger, Bianca Selzam, and Angelika Storrer argue that when modeling wordnets in OWL, it has to be decided whether to adhere to either a class model, an instance model, or a metaclass model. They discuss and compare the features of these three models by the example of the two resources GermaNet and TermNet. In several experiments, each model is assessed for its performance when querying and processing it in the context of automatic hyperlinking. Because of its compatibility with notions from traditional lexical semantics, they favor the metaclass model.

## Acknowledgement

Like its predecessor "*Linguistic Modeling of Information and Markup Languages*" [11], this book is based on contributions and numerous discussions at a corresponding conference at the *Center for Interdisciplinary Research* (*Zentrum für interdisziplinäre Forschung*, ZiF) at Bielefeld University. In the case of the present volume, this relates to the conference on "*Processing Text-Technological Resources*", organized by the research group "*Text-technological*

*Information Modeling*"[4] that was funded from 2002 to 2009 by the German Research Foundation (DFG).

The editors gratefully acknowledge financial support by the DFG and by the ZiF. We also thank the series editor and all reviewers who helped to shape this book by means of their invaluable hints. Last but not least we gratefully acknowledge all authors for their contributions to this book.

## References

[1] Abiteboul, S.: Querying semi-structured data. In: Afrati, F.N., Kolaitis, P.G. (eds.) ICDT 1997. LNCS, vol. 1186, pp. 1–18. Springer, Heidelberg (1996)

[2] Aho, A.V., Hopcroft, J.E., Ullman, J.D.: Data Structures and Algorithms. Computer Science and Information Processing, Addison-Wesley, Reading, Massachusetts (1983)

[3] Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22, 249–254 (1996)

[4] Feldman, R., Sanger, J.: The Text Mining Handbook. In: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007)

[5] Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, Mahwah (2007)

[6] Mani, I.: Automatic Summarization. John Benjamins, Amsterdam (2001)

[7] Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text 8, 243–281 (1988)

[8] Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge (2000)

[9] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)

[10] Soderland, S.: Learning information extraction rules for semi-structured and free text. Machine Learning 34(1), 233–272 (1999)

[11] Witt, A., Metzing, D. (eds.): Linguistic Modeling of Information and Markup Languages. Springer, Dordrecht (2010)

[4] `www.text-technology.de`