2nd Computational Linguistics Fall School — Institute of Cognitive Science — University of Osnabrück

Introduction to Text Technology: Using XML in Natural Language Processing

Georg Rehm

Justus-Liebig-Universität Gießen
Institut für Germanistik
Angewandte Sprachwissenschaft und Computerlinguistik
Otto-Behaghel-Straße 10 D
35394 Gießen

Telefon: +49 641 99 29052
Fax: +49 641 99 29059

Georg.Rehm@uni-giessen.de
http://www.uni-giessen.de/~g91063/
http://www.uni-giessen.de/germanistik/ascl/

September 10, 2003

Abstract

This document covers the basic structure of the course "Introduction to Text Technology: Using XML in Natural Language Processing", the topics of the individual sessions, requirements for a graded Schein, the most important websites and essential reading.

Introduction to Text Technology – Course Overview

- Week 1: 09/22/2003—09/26/2003

  1. Introduction:
     Text Technology: Definitions and Related Fields.
     A Visual Introduction to XML.
     The History of Markup Languages.
     Example of a Text Technology application.
  2. Markup Languages I:
     HTML Crash Course.
     Example: Rumelhart's Story Grammar as an XML Application.
     Introduction to the Syntax of XML-DTDs – Part 1: Elements.
  3. Markup Languages II:
     Introduction to the Syntax of XML-DTDs – Part 2: Attributes and Entities.
     Linguistic Background of Document Grammars.
     How to write Document Type Definitions.
     Parsing XML Document Instances.
     Basic XML Software: XML Parsers and XML Editors.
  4. Practical Exercise I:
     Building and parsing DTDs and XML Document Instances.
  5. Doing it with Style I:
     Cascading Style Sheets.
     Introduction to XSL.
     XPath as a Prerequisite for XSLT.

- Week 2: 09/29/2003—10/03/2003

  6. Doing it with Style II:
     XSLT.
  7. Practical Exercise II:
     Handling of a standalone XSLT processor.
     Implementation of XSLT stylesheets
  8. Related Standards:
     Beyond DTDs: XML Schema.
     XML and Related Standards.
  9. Glueing it All Together: Text Technology Applications
     Text Technology Applications.
     Programming XML interfaces: SAX and DOM.
  10. Coming Full Circle:
      The Future: The Semantic Web.
      Topics we didn't (and couldn't) cover.
      Disadvantages of XML.

Requirements for a Schein

In order to receive a graded Schein for this course, you can choose one of the following options. Regular attendance and especially practical work during the two exercises is mandatory.

1. If you already happen to

   - be in the planning or even writing phase of the final thesis for the course you are reading and if your thesis could be related to Text Technology in one way or another, or if you
   - work in an applied NLP- or Computational Linguistics research project:

   Write a paper which introduces the topic of your thesis or the project you work on and describe its potential relations to Text Technology. Where could XML-related formalisms or languages be useful? Would it be possible to implement proprietary data structures in a Text Technology way? You should make the relations as concrete as possible, i. e., you should include, e. g., example data, XSLT stylesheets, architectures, Document Type Definitions or code. The paper should not exceed 20 pages.

   Example: If your thesis deals with natural language parsing, you should discuss the representation of parse trees as XML document instances where the hierarchical structure of the parse tree is reflected by the XML structure of the instance and you could develop XML markup languages for the representation of the lexicon and the grammar.

2. If you have a background in programming:

   Develop a piece of software that utilizes Text Technology methods and write a paper as its documentation. The function of this software is open to negotiation. The paper should not exceed 20 pages.

   Example: Since the mid-90ies, corpora have become a crucial aspect of almost any NLP- or CL-related project, especially with regard to the evaluation of natural language software. Choose a language that some NL-software should process, then choose a website containing HTML documents of this language which could be processed by the abovementioned NL-software. Implement a robot and a wrapper which (a) automatically connect to this website, (b) identify webpages of interest, (c) wrap them up in simple XML code, (d) perform basic tokenization and (e) store these XML document instances in a corpus for further processing. You could even try to embed this corpus-building-process into a larger architecture, using NL-software already available at your university or department (taggers, parsers etc.). Another aspect would include the implementation of word frequency statistics, simple text summarization methods (e. g., with graphical representations based on XSLT transformations) and so on.

3. If you do not have a background in programming:

   Write a paper which discusses Text Technology methods and approaches in concrete research projects and software prototypes, which process natural language input on *one* of the following levels of language description or fields of application respectively: (a) morphology, (b) syntax/grammar, (c) semantics, (d) pragmatics, (e) natural language generation, (f) summarization, (g) web document analysis, or (h) information retrieval. Limit yourself to at most three different approaches and describe their individual advantages and disadvantages. Suggest potential improvements of the discussed approaches. The paper should not exceed 20 pages.

Important Standards, Books and Papers

Standards: World Wide Web Consortium

Adler, Sharon; Berglund, Anders; Caruso, Jeff; Deach, Stephen; Graham, Tony; Grosso, Paul; Gutentag, Eduardo; Milowski, Alex; Parnell, Scott; Richman, Jeremy and Zilles, Steve (2001): "Extensible Stylesheet Language (XSL) 1.0". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/xsl/`.

Altheim, Murray; Boumphrey, Frank; Dooley, Sam; McCarron, Shane; Schnitzenbaumer, Sebastian and Wugofski, Ted (2001): "Modularization of XHTML". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/xhtml-modularization/`.

Altheim, Murray and McCarron, Shane (2001): "XHTML 1.1 – Module-based XHTML". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/xhtml11/`.

Ayars, Jeff; Bulterman, Dick; Cohen, Aaron; Day, Ken; Hodge, Erik; Hoschka, Philipp; Hyche, Eric; Jordan, Muriel; Kim, Michelle; Kubota, Kenichi; Lanphier, Rob; Layaïda, Nabil; Michel, Thierry; Newman, Debbie; van Ossenbruggen, Jacco; Rutledge, Lloyd; Saccocio, Bridie; Schmitz, Patrick and ten Kate, Werner (2001): "Synchronized Multimedia Integration Language (SMIL 2.0)". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/smil20/`.

Baker, Mark; Ishikawa, Masayasu; Matsui, Shinichi; Stark, Peter; Wugofski, Ted and Yamakami, Toshihiko (2000): "XHTML Basic". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/xhtml-basic/`.

Berglund, Anders; Boag, Scott; Chamberlin, Don; Fernandez, Mary F.; Kay, Michael; Robie, Jonathan and Siméon, Jérôme (2002): "XML Path Language (XPath) – Version 2.0". Technische Spezifikation (Working Draft), W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/xpath20/`.

Bos, Bert; Lie, Håkon Wium; Lilley, Chris and Jacobs, Ian (1998): "Cascading Style Sheets, Level 2 – CSS2 Specification". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/REC-CSS2/`.

Box, Don; Ehnebuske, David; Kakivaya, Gopal; Layman, Andrew; Nielsen, Noah Mendelsohn Henrik Frystyk; Thatte, Satish and Winer, Dave (2000): "Simple Object Access Protocol (SOAP) 1.1". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/SOAP/`.

Bray, Tim; Hollander, Dave and Layman, Andrew (1999): "Namespaces in XML". Technische Spezifikation, W3C (World Wide Web Consortium).

Bray, Tim; Paoli, Jean; Sperberg-McQueen, C. M. and Maler, Eve (2000): "Extensible Markup Language (XML) 1.0 (Second Edition)". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/2000/REC-xml-20001006`.

Brickley, Dan and Guha, R.V. (2003): "RDF Vocabulary Description Language 1.0: RDF Schema". Technische Spezifikation (Working Draft), W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/rdf-schema/`.

Christensen, Erik; Curbera, Francisco; Meredith, Greg and Weerawarana, Sanjiva (2001): "Web Services Description Language (WSDL) 1.1". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/wsdl`.

Clark, James (1999): "XSL Transformations (Version 1.0)". Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: `http://www.w3.org/TR/xslt/`.

Clark, James and DeRose, Steve (1999): “XML Path Language (XPath) – Version 1.0”. Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: http://www.w3.org/TR/xslt/.

DeRose, Steve; Maler, Eve and Orchard, David (2001): “XML Linking Language (XLink) Version 1.0”. Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: http://www.w3.org/TR/xlink/.

Fallside, David C.; Thompson, Henry S.; Beech, David; Maloney, Murray; Mendelsohn, Noah; Biron, Paul V. and Malhotra, Ashok (2001): “XML Schema”. Technische Spezifikation, W3C (World Wide Web Consortium). W3C Recommendation. Besteht aus Part 0 (Primer), Part 1 (Structures), Part 2 (Datatypes). Online verfügbar: http://www.w3.org/XML/Schema.

Ferraiolo, Jon; Fujisawa, Jun and Jackson, Dean (2003): “Scalable Vector Graphics (SVG) 1.1 Specification”. Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: http://www.w3.org/TR/SVG11/.

Grosso, Paul; Maler, Eve; Marsh, Jonathan and Walsh, Norman (2003): “XPointer Framework”. Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: http://www.w3.org/TR/xptr-framework/.

Lassila, Ora and Swick, Ralph R. (1999): “Resource Description Framework (RDF). Model and Syntax Specification”. Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: http://www.w3.org/TR/REC-rdf-syntax/.

Mitra, Nilo; Gudgin, Martin; Hadley, Marc; Mendelsohn, Noah; Moreau, Jean-Jacques and Nielsen, Henrik Frystyk (2002): “SOAP Version 1.2”. Technische Spezifikation, W3C (World Wide Web Consortium). W3C Recommendation. Besteht aus Part 0 (Primer), Part 1 (Messaging Framework), Part 2 (Adjuncts). Online verfügbar: http://www.w3.org/2002/ws/.

Pemberton, Steven (2000): “XHTML 1.0: The Extensible Hypertext Markup Language”. Technische Spezifikation, W3C (World Wide Web Consortium).

Pemberton, Steven (2002): “XHTML 1.0: The Extensible Hypertext Markup Language (Second Edition)”. Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: http://www.w3.org/TR/xhtml1/.

Raggett, Dave; Hors, Arnaud Le and Jacbos, Ian (1999): “HTML 4.01 Specification”. Technische Spezifikation, W3C (World Wide Web Consortium). Online verfügbar: http://www.w3.org/TR/html401/.

Standards: ISO

ISO 10179 (1996): “Information Processing – Processing Languages – Document Style Semantics and Specification Language (DSSSL)”. Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: http://www.ornl.gov/sgml/wg8/.

ISO 10744 (1997): “Information Processing – Hypermedia/Time-Based Structuring Language (HyTime) – Second Edition”. Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: http://www.ornl.gov/sgml/wg8/.

ISO 15445 (2000): “ISO/IEC 15445: Information Technology – Document Description and Processing Languages – HyperText Markup Language (HTML)”. Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: http://purl.org/NET/ISO+IEC.15445/15445.html.

ISO 639 (1998): “ISO/IEC 639: Codes for the representation of names of languages. Part 1 (1988), Part 2 (1998)”. Internationaler Standard, International Organization for Standardization, Genf.

ISO 8879 (1986): “Information Processing – Text and Office Information Systems – Standard Generalized Markup Language”. Internationaler Standard, International Organization for Standardization, Genf.

ISO 8879 TC2 (1998): “ISO/IEC JTC1/SC34 N0029 – Document Description and Processing Languages – Technical Corrigendum 2 to ISO 8879:1986 (Annex K: WebSGML Adaptations; Annex L: Declaration for XML)”. Anhang/Revision eines internationalen Standards, International Organization for Standardization, Genf. Online verfügbar: http://www.sgmlsource.com/8879/n0029.htm.

ISO 9070 (1991): “ISO/IEC 9070: Information Technology – SGML Support Facilities – Registration Procedures for Public Text Owner Identifiers”. Internationaler Standard, International Organization for Standardization, Genf.

ISO/IEC 13250 (2000): “Information Technology – Document Description and Processing Languages – Topic Maps”. Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: http://www.ornl.gov/sgml/wg4/.

Standards: Requests for Comments – RFCs

RFC 2376 (1998): “XML Media Types”. Network Working Group – Request for Comments (RFC). E. James Whitehead und Makoto Murata. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2396 (1998): “Uniform Resource Identifiers (URI): Generic Syntax”. Network Working Group – Request for Comments (RFC). Tim Berners-Lee, Roy Fielding und Larry Masinter. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2413 (1999): “Dublin Core Metadata for Resource Discovery”. Network Working Group – Request for Comments (RFC). Stuart L. Weibel, John A. Kunze, Carl Lagoze und Misha Wolf. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2616 (1999): “Hypertext Transfer Protocol – HTTP/1.1”. Network Working Group – Request for Comments (RFC). Roy T. Fielding, James Gettys, Jeffrey C. Mogul, Henrik Frystyk Nielsen, Larry Masinter, Paul J. Leach und Tim Berners-Lee. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2731 (1999): “Encoding Dublin Core Metadata in HTML”. Network Working Group – Request for Comments (RFC). John A. Kunze. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2854 (2000): “The ‘text/html’ Media Type”. Network Working Group – Request for Comments (RFC). Dan Connolly und Larry Masinter. Online verfügbar: http://www.ietf.org/rfc/.

Text Technological Applications

Chen, Li-Qun; Xie, Xing; Ma, Wei-Ying; Zhang, Hong-Jiang; Zhou, Heqin and Feng, Huanqing (2003): “DRESS: A Slicing Tree Based Web Page Representation for Various Display Sizes”. In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Coling (1994): *COLING 94 – The 15th International Conference on Computational Linguistics*, Kyoto, Japan. Association for Computational Linguistics.

DCES (2003): “Dublin Core Metadata Element Set, Version 1.1: Reference Description”. Dublin Core Metadata Initiative, DCMI. Online verfügbar: http://dublincore.org/documents/2003/02/04/dces/.

DCMT (2003): “DCMI Metadata Terms”. Dublin Core Metadata Initiative, DCMI. Online verfügbar: http://dublincore.org/documents/dcmi-terms/.

DCTV (2003): “DCMI Type Vocabulary”. Dublin Core Metadata Initiative, DCMI. Online verfügbar: http://dublincore.org/documents/dcmi-type-vocabulary/.

Dunlop, Dominic (1995): “Practical Considerations in the Use of TEI Headers in a Large Corpus”. *Computers and the Humanities* (29): pp. 85–98.

Fujii, Atsushi and Ishikawa, Tetsuya (2000): “Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-Structured Texts”. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2000)*. pp. 488–495.

Fujii, Atsushi and Ishikawa, Tetsuya (2001): “Organizing Encyclopedic Knowledge based on the Web and its Application to Question Answering”. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)*. pp. 196–203.

Gupta, Suhit; Kaiser, Gail; Neistadt, David and Grimm, Peter (2003): “DOM-based Content Extraction of HTML Documents”. In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Ide, Nancy (1994): “Encoding Standards for Large Text Resources: The Text Encoding Initiative”. In: Coling (1994), pp. 574–578.

Kim, Sun and Zhang, Byoung-Tak (2000): “Web-Document Retrieval by Genetic Learning of Importance Factors for HTML Tags”. In: *Proceedings of the International Workshop on Text and Web Mining (PRICAI 2000)*, edited by Tan, Ah-Hwee and Yu, Philip S. Melbourne, pp. 13–23.

Klabunde, Ralf; Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Langer, Hagen and Schiehlen, Michael (editors) (2003): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg: Spektrum, 2nd edition. Erscheint.

Labrou, Yannis and Finin, Tim (1999): “Yahoo! as an Ontology – Using Yahoo! Categories to Describe Documents”. In: *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM ’99)*. ACM Press, pp. 180–187.

Leech, Geoffrey; Garside, Roger and Bryant, Michael (1994): “CLAWS4: The Tagging of the British National Corpus”. In: Coling (1994), pp. 622–628.

Lobin, Henning (1998): “Intelligente Dokumente – Inhaltliche Strukturierung und flexible Hypertextualisierung von Multimedia-Dokumenten”. In: *Forschung an der Universität Bielefeld*, Bielefeld: Universität Bielefeld, pp. 35–40.

Lobin, Henning (1999a): “Intelligente Dokumente – Linguistische Repräsentation komplexer Inhalte für die hypermediale Wissensvermittlung”. In: Lobin (1999b), pp. 155–178.

Lobin, Henning (editor) (1999b): *Text im digitalen Medium – Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering*. Wiesbaden: Westdeutscher Verlag.

Lobin, Henning (2000): *Informationsmodellierung in XML und SGML*. Berlin, Heidelberg, New York etc.: Springer.

Lobin, Henning (editor) (2001): *Sprach- und Texttechnologie in digitalen Medien – Proceedings der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung*, Justus-Liebig-Universität Gie"sen. Gesellschaft für linguistische Datenverarbeitung.

Lobin, Henning and Lemnitzer, Lothar (editors) (2003): *Texttechnologie – Anwendungen und Perspektiven*. Tübingen: Stauffenburg. Erscheint.

Lobin, Henning and Stührenberg, Maik (2003): “XML-strukturierte Learning Objects”. In: *Sprache zwischen Theorie und Technologie – Festschrift für Wolf Papproté zum 60. Geburtstag*, edited by Cyrus, Lea; Feddes, Hendrik; Schumacher, Frank and Steiner, Petra, Wiesbaden: Deutscher Universitäts-Verlag, Sprachwissenschaft, pp. 185–198.

Maler, Eve and Andaloussi, Jeanne El (1996): *Developing SGML DTDs – From Text to Model to Markup*. Upper Saddle River: Prentice Hall.

Martelli, Silvia; Carroll, Jeremy J. and Signore, Oreste (2003): "Syntax for Semantic Enriching of Web Pages". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Martin, Philippe and Eklund, Peter (1999): "Embedding Knowledge in Web Documents". *Computer Networks* (31): pp. 1403–1419.

McKelvie, David; Brew, Chris and Thompson, Henry (1997): "Using SGML as a Basis for Data-Intensive NLP". In: *Proceedings of Applied Natural Language Processing (ANLP) 97*. Association for Computational Linguistics, Washington D. C. Online verfügbar: `http://www.ltg.hcrc.ed.ac.uk/~dmck/anlp97.ps`.

Mehler, Alexander and Lobin, Henning (editors) (2003): *Automatische Textanalyse – Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*. Wiesbaden: Westdeutscher Verlag. Erscheint.

Mikheev, Andrei; Grover, Claire and Moens, Marc (1999): "XML Tools and Architecture for Named Entity Recognition". *Markup Languages* 1 (3): pp. 89–113.

Minning, Holger; Puskás, Csilla and Salisbury, Justin (2002): "Rhetorisches Parsing deutschsprachiger Texte". In: *Proceedings of the 12th Student Conference on Computational Linguistics (TaCoS 2002)*, edited by Reitter, David. Potsdam. Online verfügbar: `http://www.ling.uni-potsdam.de/tacos/proceedings/`.

Möhr, Wiebke and Schmidt, Ingrid (editors) (1999): *SGML und XML – Anwendungen und Perspektiven*. Berlin, Heidelberg, New York etc.: Springer.

Myllymaki, Jussi (2001): "Effective Web Data Extraction with Standard XML Technologies". In: *Proceedings of the 10th International World Wide Web Conference (WWW-10)*. Hong Kong, pp. 689–696.

Nagao, Katashi and Hasida, Kôiti (1998): "Automatic Text Summarization Based on the Global Document Annotation". In: *COLING 98 – The 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Association for Computational Linguistics, Montreal, Quebec, Kanada, volume 2, pp. 917–921. 2 Bände.

Pepper, Steve and Moore, Graham (2001): "XML Topic Maps (XTM) 1.0". Technische Spezifikation, TopicMaps.Org. Online verfügbar: `http://www.topicmaps.org/xtm/1.0/`.

Potok, Thomas E.; Elmore, Mark T.; Reed, Joel W. and Samatova, Nagiza F. (2002): "An Ontology-based HTML to XML Conversion Using Intelligent Agents". In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii: IEEE Computer Society.

Rehm, Georg (1998): *Vorüberlegungen zur automatischen Zusammenfassung deutschsprachiger Texte mittels einer SGML- und DSSSL-basierten Repräsentation von RST-Relationen*. Magisterarbeit, Studiengang Computerlinguistik und Künstliche Intelligenz, Universität Osnabrück. Online verfügbar: `http://www.uni-giessen.de/~g91063/papers.shtml`.

Rehm, Georg (2003a): "Das World Wide Web". In: Klabunde et al. (2003). Erscheint.

Rehm, Georg (2003b): "Texttechnologie und das World Wide Web – Anwendungen und Perspektiven". In: Lobin and Lemnitzer (2003), pp. 433–464. Erscheint.

Rehm, Georg (2003c): "Texttechnologische Grundlagen". In: Klabunde et al. (2003). Erscheint.

Rehm, Georg and Reinsch, Markus (2001): "Die Chronik der "'Chronik"' – Über die Konvertierung und Weiterverarbeitung proprietär annotierter Daten". In: *Sprach- und Texttechnologie in digitalen Medien – Proceedings der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung*, edited by Lobin, Henning. Gesellschaft für linguistische Datenverarbeitung, pp. 211–221.

Sahuguet, Arnaud and Azavant, Fabien (1999): "Looking at the Web through XML Glasses". In: *Proceedings of the 4th International Conference on Cooperative Information Systems (CoopIS '99)*. Edinburgh: IEEE Computer Society Press, pp. 148–159.

Seo, Heekyoung and Choi, Jaeyoung Yang Joongmin (2001): "Knowledge-based Wrapper Generation by Using XML". In: *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*. Seattle.

Sigletos, Georgios; Farmakiotou, Dimitra; Stamatakis, Kostas; Paliouras, Georgios and Karkaletsis, Vangelis (2003): "Annotating Web pages for the needs of Web Information Extraction applications". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Sperberg-McQueen, C. M. and Burnard, Lou (editors) (2002): *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium; Humanities Computing Unit, University of Oxford.

Sperberg-McQueen, C. M. and Goldstein, Robert F. (1994): "HTML to the Max – A Manifesto for Adding SGML Intelligence to the World-Wide Web". In: *Proceedings of the Second International WWW Conference – Mosaic and the Web*. Chicago. Online verfügbar: `http://www.uic.edu/~cmsmcq/htmlmax.html`.

Thierry, Declerck and Wittenburg, Peter (2001): "Xml and NLP: Their Interaction and their Role for HLT Applications". In: *Proceedings of the First NLP and XML Workshop*. Tokyo.

Ule, Tylman and Hinrichs, Erhard (2003): "Linguistische Annotation". In: Lobin and Lemnitzer (2003), pp. 217–243. Erscheint.

Ule, Tylman and Müller, Frank Henrik (2003): "KaRoPars: Ein System zur linguistischen Annotation großer Text-Korpora des Deutschen". In: Mehler and Lobin (2003). Erscheint.

Volk, Martin (1998): "Markup of a Test Suite with SGML". In: *Linguistic Databases*, edited by Nerbonne, John, Cambridge: Cambridge University Press, number 77 in CSLI Lecture Notes, pp. 59–76. Online verfügbar: `http://www.ifi.unizh.ch/CL/volk/papers/SGMLMarkup.ps.gz`.

Volk, Martin (2000): "Scaling up: Using the WWW to resolve PP attachments and ambiguities". In: *KONVENS-2000 / Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache*, edited by Zühlke, Werner and Schukat-Talamazzini, Ernst Günter. Ilmenau: VDE Verlag, pp. 151–155.

Volk, Martin (2001): "Exploiting the WWW as a corpus to resolve PP attachment ambiguities". In: *Proceedings of the Corpus Linguistics 2001 Conference*, edited by Rayson, Paul; Wilson, Andrew; McEnery, Tony; Hardie, Andrew and Khoja, Shereen. Lancaster, pp. 601–606.

Volk, Martin (2002): "Using the Web as Corpus for Linguistic Research". In: *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, Publications of the Department of General Linguistics 3, University of Tartu. Online verfügbar: `http://www.ifi.unizh.ch/CL/volk/publications.html`.

Witt, Andreas (1999): "SGML und Linguistik". In: Lobin (1999b), pp. 121–154.

Witt, Andreas (2003): "Linguistische Informationsmodellierung mit XML". In: Mehler and Lobin (2003). Erscheint.

Wolff, Christian (2003): "Systemarchitekturen – Aufbau texttechnologischer Anwendungen". In: Lobin and Lemnitzer (2003), pp. 165–192. Erscheint.

Corpora and the Web

Campos, João and Silva, Mário J. (2001): "Versus: A Model for a Web Repository". In: *CRC'01 – 4ª Conferência de Redes de Computadores*. Covilhã.

Chung, Christina Yip; Gertz, Michael and Sundaresan, Neel (2001): "Quixote: Building XML Repositories from Topic Specific Web Documents". In: *Fourth International Workshop on the Web and Databases (WebDB 2001)*, edited by Giansalvatore Mecca, Jérôme Siméon. Santa Barbara, pp. 103–108.

Cowie, Jim; Ludovik, Evgeny and Zacharski, Ron (1998): "An Autonomous, Web-based, Multilingual Corpus Collection Tool". In: *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*. Moncton, pp. 142–148. Online verfügbar: `http://crl.nmsu.edu/~raz/langrec/nlpia.htm`.

Dewe, Johan; Karlgren, Jussi and Bretan, Ivan (1998): "Assembling a Balanced Corpus from the Internet". In: *Proceedings of the 11th Nordic Conference of Computational Linguistics*. Kopenhagen, pp. 100–107.

Fairon, Cédrick (2000): "GlossaNet: Parsing a Web Site as a Corpus". *Linguisticæ Investigationes* 22 (2): pp. 327–340.

Fletcher, William H. (2001): "Concordancing the Web with KWICFinder". In: *Proceedings of the 3rd North American Symposium on Corpus Linguistics and Language Teaching*. Boston.

Grefenstette, Gregory (1999): "The World Wide Web as a resource for example-based machine translation tasks". In: *Proceedings of the ASLIB Conference on Translating and the Computer*. London.

Hirai, Jun; Raghavan, Sriram; Garcia-Molina, Hector and Paepcke, Andreas (2000): "WebBase: A Repository of Web Pages". In: *Proceedings of the 9th International World Wide Web Conference*. Amsterdam, pp. 277–293.

Ide, Nancy and Véronis, Jean (1994): "MULTEXT: Multilingual Text Tools and Corpora". In: *COLING 94 – The 15th International Conference on Computational Linguistics*. Association for Computational Linguistics, Kyoto, Japan, volume 1, pp. 588–592.

Jones, Rosie and Ghani, Rayid (2000): "Automatically Building a Corpus for a Minority Language from the Web". In: *Proceedings of the Student Workshop at the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*. Hong Kong. Online verfügbar: `http://www.cs.cmu.edu/~webkb/`.

Kehoe, Andrew and Renouf, Antoinette (2002): "WebCorp: Applying the Web to Linguistics and Linguistics to the Web". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Kilgarriff, Adam (2001): "Web as Corpus". In: *Proceedings of the Corpus Linguistics 2001 Conference*, edited by Rayson, Paul; Wilson, Andrew; McEnery, Tony; Hardie, Andrew and Khoja, Shereen. Lancaster, pp. 342–344.

Li, Fang; Sheng, Huanye and Weisweber, Wilhelm (2001): "World Wide Web – A Multilingual Language Resource". In: *Web Intelligence: Research and Development*, edited by Zhong, Ning; Yao, Yiyu; Liu, Jiming and Ohsuga, Setsuo. Berlin, Heidelberg, New York etc.: Springer, number 2198 in Lecture Notes in Artificial Intelligence, pp. 373–378.

Morley, Barry; Renouf, Antoinette and Kehoe, Andrew (2003): "Linguistic Research with XML/RDF-aware WebCorp tool". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Rehm, Georg (2001): "korpus.html – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten". In: *Proceedings of the GLDV Spring Meeting 2001*, edited by Lobin, Henning. Gesellschaft für linguistische Datenverarbeitung (Society for Computational Linguistics and Language Technology), Giessen, Germany, pp. 93–103. Online verfügbar: http://www.uni-giessen.de/fb09/ascl/gldv2001/.

Resnik, Philip and Smith, Noah A. (2002): "The Web as a Parallel Corpus". Technical Report UMIACS-TR-2002-61, University of Maryland. Online verfügbar: http://www.umiacs.umd.edu/~resnik/pubs.html.

Language Identification and the Web

Grefenstette, Gregory and Nioche, Julien (2000): "Estimation of English and non-English Language Use on the WWW". In: *Proceedings of RIAO'2000: Content-Based Multimedia Information Access*. Paris, pp. 237–246. Online verfügbar: http://de.arXiv.org/abs/cs.CL/0006032.

Langer, Stefan (2002): "Grenzen der Sprachenidentifizierung". In: *KONVENS 2002 – 6. Konferenz zur Verarbeitung natürlicher Sprache*, edited by Busemann, Stefan. Saarbrücken: Deutsches Forschungszentrum für Künstliche Intelligenz, pp. 99–106. DFKI Document D-02-01.

Lavoie, Brian F. and O'Neill, Edward T. (1999): "How "World Wide" is the Web? Trends in the Internationalization of Web Sites". In: *Annual Review of OCLC Research 1999*, Dublin: OCLC Online Computer Library Center. Online verfügbar: http://www.oclc.org.

Muthusamy, Yeshwant K. and Spitz, Lawrence (1998): "Automatic Language Identification". In: *Survey of the State of the Art in Human Language Technology*, edited by Cole, Ronald; Mariani, Joseph; Uszkoreit, Hans; Varile, Giovanni Battista; Zaenen, Annie and Zampolli, Antonio, Cambridge: Cambridge University Press, pp. 314–317. Online verfügbar: http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Rehm, Georg (2003): "Ontologie-basierte Hypertextsorten-Klassifikation". In: *Automatische Textanalyse – Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, edited by Mehler, Alexander and Lobin, Henning, Wiesbaden: Westdeutscher Verlag. Erscheint.

Crawling the Web

Chakrabarti, Soumen; van den Berg, Martin and Dom, Byron (1999): "Focused crawling: A new approach to topic-specific Web resource discovery". *Computer Networks* 31 (11–16): pp. 1623–1640.

Cowie, Jim; Ludovik, Evgeny and Zacharski, Ron (1998): "An Autonomous, Web-based, Multilingual Corpus Collection Tool". In: *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*. Moncton, pp. 142–148. Online verfügbar: http://crl.nmsu.edu/~raz/langrec/nlpia.htm.

Heydon, Allan and Najork, Marc (1999): "Mercator: A Scalable, Extensible Web Crawler". *World Wide Web* 2 (4): pp. 219–229.

O'Neill, Edward T.; McClain, Patrick D. and Lavoie, Brian F. (1997): "A Methodology for Sampling the World Wide Web". In: *Annual Review of OCLC Research 1997*, Dublin: OCLC Online Computer Library Center. Online verfügbar: http://www.oclc.org.

Raghavan, Sriram and Garcia-Molina, Hector (2001): "Crawling the Hidden Web". In: *Proceedings of the 27th International Conference on Very Large Databases (VLDB)*. pp. 129–138.

Rehm, Georg (2001): "korpus.html – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten". In: *Proceedings of the GLDV Spring Meeting 2001*, edited by Lobin, Henning. Gesellschaft für linguistische Datenverarbeitung (Society for Computational Linguistics and Language Technology), Giessen, Germany, pp. 93–103. Online verfügbar: http://www.uni-giessen.de/fb09/ascl/gldv2001/.

Thelwall, Mike (2001): "A Web Crawler Design for Data Mining". *Journal of Information Science* 27 (5): pp. 319–326.

Turau, Volker (1998): "Web-Roboter". *Informatik Spektrum* 21 (3): pp. 159–160.

The Semantic Web

Berners-Lee, Tim (1999): *Weaving the Web – The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: Harper San Francisco.

Berners-Lee, Tim; Cailliau, Robert; Groff, Jean-François and Pollermann, Bernd (1992): "World-Wide Web: The Information Universe". *Electronic Networking: Research, Applications and Policy* 1 (2): pp. 52–58.

Berners-Lee, Tim; Hendler, James and Lassila, Ora (2001): "The Semantic Web". *Scientific American* 284 (5): pp. 34–43.

Dill, Stephen; Eiron, Nadav; Gibson, David; Gruhl, Daniel; Guha, Ramanathan; Jhingran, Anant; Kanungo, Tapas; Rajagopalan, Sridhar; Tomkins, Andrew; Tomlin, John A. and Zien, Jason Y. (2003): "SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Doan, AnHai; Madhavan, Jayant; Domingos, Pedro and Halevy, Alon (2002): "Learning to Map between Ontologies on the Semantic Web". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Klein, Michael; Fensel, Dieter; van Harmelen, Frank and Horrocks, Ian (2001): "The Relation Between Ontologies and XML Schemas". *Linköping Electronic Articles in Computer and Information Science* 6 (4). Online verfügbar: http://www.ida.liu.se/ext/epa/cis/.

Web Document Analysis

Asirvatham, Arul Prakash and Ravi, Kranthi Kumar (2001): "Web Page Classification Based on Document Structure". Technical Report, International Institute of Information Technology, Hyderabad. Online verfügbar: http://www.iiit.net/stud_pub.htm.

Buyukkokten, Orkut; Garcia-Molina, Hector; Paepcke, Andreas and Winograd, Terry (2000): "Power Browser: Efficient Web Browsing for PDAs". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI'00)*. The Hague.

Carchiolo, Vincenza; Longheu, Alessandro and Malgeri, Michele (2002): "Extraction of Hidden Semantics from Web Pages". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2002*, edited by Yin, Hujun; Allinson, Nigel; Freeman, Richard; Keane, John and Hubbard, Simon, Berlin, Heidelberg, New York etc.: Springer, number 2412 in Lecture Notes in Computer Science, pp. 117–122.

Chan, Michael and Yu, Gin (1999): "Extracting Web Design Knowledge: The Web De-Compiler". In: *IEEE International Conference on Multimedia Computing and Systems (ICMCS 1999)*. IEEE Computer Society, Florence, volume 2, pp. 547–552.

Chen, Yu; Ma, Wei-Ying and Zhang, Hong-Jiang (2003): "Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Cutler, Michael; Shih, Yungming and Meng, Weiyi (1997): "Using the Structure of HTML Documents to Improve Retrieval". In: *USENIX Symposium on Internet Technologies and Systems (NSITS '97)*. Monterey, pp. 241–251.

DiPasquo, Dan (1998): "Using HTML Formatting to Aid in Natural Language Processing on the World Wide Web". Senior Honors Thesis, School of Computer Science, Carnegie Mellon University. Online verfügbar: `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/`.

Dorai, Gautham K. and Yacoob, Yaser (2002): "Embedded Grammar Tags: Advancing Natural Language Interaction on the Web". *IEEE Intelligent Systems* 17 (1): pp. 48–53.

Jiang, Yuan (1998): *Record-Boundary Detection in Web Documents*. Master's thesis, Brigham Young University.

Rehm, Georg (2003): "Hypertextsorten-Klassifikation als Grundlage generischer Informationsextraktion". In: *Automatische Textanalyse – Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, edited by Mehler, Alexander and Lobin, Henning, Wiesbaden: Westdeutscher Verlag. Erscheint.

Sahuguet, Arnaud and Azavant, Fabien (1999): "Looking at the Web through XML Glasses". In: *Proceedings of the 4th International Conference on Cooperative Information Systems (CoopIS '99)*. Edinburgh: IEEE Computer Society Press, pp. 148–159.

Walker, Derek (1999): "Taking Snapshots of the Web with a TEI Camera". *Computers and the Humanities* 33 (1–2): pp. 185–192.

Wrapping and Information Extraction

Abiteboul, Serge; Buneman, Peter and Suciu, Dan (2000): *Data on the Web – From Relations to Semistructured Data and XML*. San Francisco: Morgan Kaufmann.

Adelberg, Brad (1998): "NoDoSE – A Tool for Semi-Automatically Extracting Structured and Semistructured Data from Text Documents". In: *Proceedings of the 1998 ACM International Conference on Management of Data (SIGMOD'98)*. Seattle, pp. 283–294.

Adelberg, Brad and Denny, Matt (1999): "Building Robust Wrappers for Text Sources". Technical Report, Department of Computer Science, Northwestern University.

Baumgartner, Robert; Flesca, Sergio and Gottlob, Georg (2001): "Visual Web Information Extraction with Lixto". In: *Proceedings of the 27th VLDB Conference*. Rom, pp. 119–128.

Bollacker, Kurt; Lawrence, Steve and Giles, C. Lee (1998): "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications". In: *Proceedings of the Second International Conference on Autonomous Agents*, edited by Sycara, Katia P. and Wooldridge, Michael. New York: ACM Press, pp. 116–123.

Buttler, David; Liu, Ling and Pu, Calton (2001): "A Fully Automated Object Extraction System for the World Wide Web". In: *Proceedings of the 2001 International Conference on Distrubuted Computing Systems (ICDCS '01)*. Phoenix.

Chakrabarti, Soumen (2003): *Mining the Web – Discovering Knowledge from Hypertext Data*. Amsterdam, Boston, London etc.: Morgan Kaufmann.

Eikvil, Line (1999): "Information Extraction from World Wide Web – A Survey". Technical Report 945, Norwegian Computing Center.

Gao, Xiaoying and Sterling, Leon (1999): "AutoWrapper: Automatic Wrapper Generation for Multiple Online Services". In: *Proceedings of Asia Pacific Web Conference 1999 (APWeb99)*. pp. 61–70.

Gupta, Suhit; Kaiser, Gail; Neistadt, David and Grimm, Peter (2003): "DOM-based Content Extraction of HTML Documents". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Liu, Ling; Pu, Calton and Han, Wei (2000): "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources". In: *Proceedings of the International Conference on Data Engineering (ICDE)*. pp. 611–621.

Myllymaki, Jussi (2001): "Effective Web Data Extraction with Standard XML Technologies". In: *Proceedings of the 10th International World Wide Web Conference (WWW-10)*. Hong Kong, pp. 689–696.

Paradis, François (2000): "Information Extraction and Gathering for Search Engines: The Taylor Approach". RIAO (Recherche d'Informations Assistée par Ordinateur), Paris, France.

Seo, Heekyoung and Choi, Jaeyoung Yang Joongmin (2001): "Knowledge-based Wrapper Generation by Using XML". In: *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*. Seattle.

Soderland, Stephen (1997): "Learning to Extract Text-Based Information from the World Wide Web". In: *Proceedings of the 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-97)*, edited by Heckerman, David; Mannila, Heikki and Pregibon, Daryl. Newport Beach: AAAI Press, pp. 251–254.

Searching the Web

Brin, Sergey and Page, Lawrence (1998): "The Anatomy of a Large-Scale Hypertextual Web Search Engine". *Computer Networks and ISDN Systems* 30 (1–7): pp. 107–117.

Chakrabarti, Soumen; Dom, Byron; Kumar, S. Ravi; Raghavan, Prabhakar; Rajagopalan, Sridhar; Tomkins, Andrew; Kleinberg, Jon M. and Gibson, David (1999): "Hypersearching the Web". *Scientific American* 280 (6): pp. 54–60.

Hu, Wen-Chen; Chen, Yining; Schmalz, Mark S. and Ritter, Gerhard X. (2001): "An Overview of World Wide Web Search Technologies". In: *Proceedings of the 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001)*. Orlando.

Huang, Lan (2000): "A Survey on Web Information Retrieval Technologies". Technical Report, Experimental Computer Systems Laboratory (ECSL), State University of New York (SUNY). Online verfügbar: `http://www.cs.sunysb.edu/~lanhuang/`.

Automatic Summarization in the Web

Buyukkokten, Orkut; Garcia-Molina, Hector and Paepcke, Andreas (2001a): "Accordion summarization for end-game browsing on PDAs and cellular phones". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, pp. 213–220.

Buyukkokten, Orkut; Garcia-Molina, Hector and Paepcke, Andreas (2001b): "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices". In: *Proceedings of the 10th International World Wide Web Conference (WWW 2001)*. Hong Kong.

Corston-Oliver, Simon (1998): *Computing Representations of the Structure of Written Discourse*. Ph.D. thesis, University of California, Santa Barbara. Zugleich: Microsoft Research Technical Report, MSR-TR-98-15.

Leong, H.; Kapur, S. and de Vel, O. (1997): "Text Summarisation for Knowledge Filtering Agents in Distributed Heterogeneous Environments". In: Mahesh (1997b), pp. 87–94. Online verfügbar: http://crl.nmsu.edu/users/mahesh/aaai-web-nlp-symposium/proceedings.html.

Lok, Simon and Kan, Min-Yen (2003): "Employing Natural Language Summarization and Automated Layout for Effective Presentation and Navigation of Information Retrieval Results". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Mahesh, Kavi (1997a): "Hypertext Summary Extraction for Fast Document Browsing". In: Mahesh (1997b), pp. 95–103. Online verfügbar: http://crl.nmsu.edu/users/mahesh/aaai-web-nlp-symposium/proceedings.html.

Mahesh, Kavi (editor) (1997b): *Working Notes of the AAAI Spring Symposium: Natural Language Processing for the World Wide Web*, Stanford. American Association for Artificial Intelligence. Online verfügbar: http://crl.nmsu.edu/users/mahesh/aaai-web-nlp-symposium/proceedings.html.

Rehm, Georg (1998): *Vorüberlegungen zur automatischen Zusammenfassung deutschsprachiger Texte mittels einer SGML- und DSSSL-basierten Repräsentation von RST-Relationen*. Magisterarbeit, Studiengang Computerlinguistik und Künstliche Intelligenz, Universität Osnabrück. Online verfügbar: http://www.uni-giessen.de/~g91063/papers.shtml.

Natural language processing helps computer to understand human language as it is spoken.Real world use of natural languages such as English,Hindi,German,French etc doesn't have a formulatedâ€¦Â Over the years there have been many advancements in Natural language processing. NLP Terminologies : Lets understand the Basic Terminologies used in NLP : Tokenization,Corpus or Corpora,Stemming,Bag of Words,Stop Words,Tf-idf,Disambiguation,Topic Models ,Word Boundaries. Tokenization : Tokenization is a process to split longer strings into smaller pieces.Large documents can be tokenized into paragraphs,Paragraphs can be tokenized into sentences and sentences can be tokenized into phrases,words or letters. Corpus or Corpora Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit by Steven Bird Paperback $38.99. In Stock. Ships from and sold by Amazon.com.Â "Instructors looking for a good introductory text to use in a course devoted to computational linguistics should consider this book as a strong candidate if they wish to emphasize a linguistics approach. The book is interestingly written with many insightful discussions, and it is the only (introductory) computational linguistics textbook that looks at the field from a linguist's point of view." NLTK (Natural Language Toolkit) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to many corpora and lexical resources. Also, it contains a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Best of all, NLTK is a free, open source, community-driven project. Weâ€™ll use this toolkit to show some basics of the natural language processing field. For the examples below, Iâ€™ll assume that we have imported the NLTK toolkit. We can do this like this: import nltk.