1 Das World Wide Web

Georg Rehm

Im Frühjahr 2003 indexiert die Suchmaschine Google (`http://www.google.com`) etwa drei Milliarden World Wide Web-Dokumente, und das Internet Archive (`http://www.archive.org`) bietet den Zugriff auf 10 Milliarden Webseiten an, die seit dem Start dieses Projekts im Jahr 1996 von der „Wayback Machine“ archiviert worden sind. Jedoch können diese Angaben nur als die Spitze des Eisberges gelten: Das „deep web“ – d. h. das gesamte Web einschließlich der von Suchmaschinen derzeit nicht erfassten bzw. nicht automatisch traversierbaren Bereiche – umfasst die Zahl von 500 Milliarden Dokumenten (*Bergman* 2000).

Für die Computerlinguistik ist das World Wide Web in dreierlei Hinsicht von Bedeutung. Zunächst wird das Web natürlich im Rahmen zahlreicher Projekte als Informationsmedium eingesetzt, um sowohl intern als auch extern Publikationen, Daten und Ressourcen anzubieten. Darüber hinaus umfasst das Web einen Textbestand kaum vorstellbarer Größe, der Dokumente unterschiedlichster Text- bzw. Hypertextsorten enthält, die in den verschiedensten Sprachen und Dialekten verfasst worden sind. Diese unmittelbar zugängliche und maschinell verarbeitbare Ressource wird z. B. zur automatischen Erstellung von Kollektionen im Bereich der Korpuslinguistik eingesetzt, in der Lexikographie zur Entdeckung von Neologismen oder zum maschinellen Aufbau von semantischen Netzen, d. h. zur Wissensakquisition. Der dritte Aspekt betrifft die Entwicklung computerlinguistischer Verfahren, mit deren Hilfe die Navigations- und Recherchefunktionen von Suchmaschinen erleichtert werden sollen, etwa durch Methoden der Informationsextraktion zur automatischen Erzeugung von Metadaten. Diese drei Anwendungsbereiche hängen unmittelbar mit einem interdisziplinären Forschungsbereich zusammen, der als „Texttechnologie“ bezeichnet und in dem gleichnamigen Unterkapitel **??** genauer vorgestellt wird. Abschnitt 1.1 geht zunächst auf die technologischen Grundlagen des WWW ein, woraufhin sich die einzelnen Unterabschnitte von 1.2 dem Einsatz des World Wide Web als Ressource in sprachverarbeitenden Systemen bzw. der Anwendung computerlinguistischer Verfahren auf Webdokumente widmen.

1.1 Technologische Grundlagen

In informationstechnologischer Hinsicht sind die drei Grundpfeiler des verteilten Hypertextsystems World Wide Web (WWW, *Berners-Lee et al.* 1992) die Textauszeichnungssprache HTML (*Hypertext Markup Language*, vgl. Unterkapitel **??**, *Texttechnologie*) zur Strukturierung und Gestaltung von Dokumenten, das Kommunikationsprotokoll HTTP (*Hypertext Transfer Protocol*, spezifiziert in RFC 2616) sowie die Möglichkeit, Dokumente durch die Angabe einer URL bzw. eines URI (*Uniform Resource Locator* bzw. *Identifier*, RFC 2396) adressieren und miteinander verknüpfen zu können. Das World Wide Web basiert auf dem Client-Server-Paradigma: Ein Server bietet einen Dienst – etwa elek-

tronische Post oder FTP (*File Transfer Protocol*) – an, und ein oder mehrere Clients greifen auf den Server zu, um diesen Dienst zu benutzen. Im Falle des World Wide Web wird der Server auch als Webserver bezeichnet, der Client als Browser oder Webbrowser. Sobald der Client auf eine URL – die Adresse, die ein Dokument oder einen Dienst eindeutig im Netz identifiziert – zugreift, findet eine Kommunikation zwischen Client und Server auf der Basis von HTTP statt, deren Ergebnis im Regelfall die Übertragung des entfernten HTML-Dokuments sowie der referenzierten, zusätzlich benötigten Dateien (z. B. Grafiken, Animationen, Style Sheets etc.) auf den lokalen Rechner sowie dessen anschließende Darstellung (Rendering) ist. Mittels Verknüpfungen (auch: Verweise, Links, Hyperlinks), die in ein HTML-Dokument integriert sind, kann der Benutzer weitere Informationsangebote anfordern.

1.2 Das WWW als computerlinguistische Ressource

Dieser Abschnitt geht auf unterschiedliche Anwendungskontexte ein, die sich auf das World Wide Web aus Sicht der Sprachtechnologie beziehen. Hierzu gehört zunächst der eigentliche Zweck des WWW als Medium zur effizienten Informationsverteilung (Abschnitt 1.2.1), die immer häufiger mit texttechnologischen Verfahren durchgeführt wird. Weiterhin werden Webdokumente, mit oder ohne enthaltenem HTML-Markup, von sprachtechnologischen Systemen – z. B. zu Testzwecken – verarbeitet (Abschnitt 1.2.2). Abschnitt 1.2.3 geht auf die Erzeugung von Korpora ein, die aus Webseiten bestehen, woraufhin Abschnitt 1.2.4 auf die Annotation linguistischer Informationen eingeht. Abschnitt 1.2.5 stellt schließlich drei konkrete sprachtechnologische Anwendungsszenarien vor.

1.2.1 Veröffentlichung von Informationen

Die in Unterkapitel **??** dargestellten Grundlagen der Texttechnologie werden aufgrund des identischen Ursprungs der zugrunde liegenden Auszeichnungssprachen auch zu Publikationszwecken im World Wide Web eingesetzt. Bereits mit Hilfe XML-basierter Markup-Sprachen (*Bray et al.* 2000) annotierte Korpora können beispielsweise zur Visualisierung oder für einen möglichst effizienten Zugriff in statische oder dynamische (X)HTML-Dokumente konvertiert werden (*Lubell* 1999). Ein weiteres Anwendungsszenario, das insbesondere auf dem Prinzip des *Single Source Publishing* beruht (vgl. Unterkapitel **??**, *Texttechnologie*), betrifft den Bereich des eLearning. Dabei werden unterschiedlich granulare Lernobjekte bei maximaler Modularität der zu vermittelnden Inhalte mit internationalen Standards (z. B. LOM, *Learning Object Metadata*) annotiert und mit zahlreichen Metadaten angereichert, so dass die Implementierung adaptiver eLearning-Umgebungen ermöglicht wird. Auf diese Weise können Lernenden, abhängig von ihrem jeweiligen Wissensstand, personalisierte Lerninhalte über das World Wide Web präsentiert werden, wodurch auch die automatische Generierung druckbarer Unterlagen ermöglicht wird (vgl. Unterkapitel **??**, *Sprachlehr- und lernsysteme* sowie *Lobin und Stührenberg* 2003).

1.2.2 Webseiten und computerlinguistische Methoden

Neben der Publikation von Inhalten im WWW werden Webdokumente auch häufig benutzt, um mit den enthaltenen textuellen Daten computerlinguistische Systeme zu evaluieren (z. B. Tagger, Parser, Textkategorisierungsverfahren und Information-Retrieval-Algorithmen). Hierzu werden häufig Korpora von Webseiten aufgebaut (vgl. Unterabschnitt 1.2.3), aus denen vor der Durchführung der Tests meist das enthaltene HTML-Markup entfernt wird.

Man kann sprachverarbeitende Systeme aber auch mit Informationen *über* HTML versorgen, um zusätzliche Daten in den Analyseprozess einfließen zu lassen, z. B. dass Einbettungen von Phrasen in HTML-Elemente wie `<strong>` oder `<em>` wichtige Informationen darstellen oder dass ein als Überschrift (`<h1>` bis `<h6>`) markierter Text vermutlich kein vollständiger Satz sondern lediglich eine Nominalphrase sein dürfte, um daraufhin eingeschränkte Parsing-Verfahren zur Analyse von NPs anzuwenden (*Kunze und Rösner* 2003). Neben der textuellen Information ist diese metasprachliche Ebene für die Unterstützung des automatischen Analyseprozesses von immenser Bedeutung und wird sich in künftigen Theorien und Systemen – insbesondere im Bereich der Suchmaschinen und des Web Mining (*Chakrabarti* 2003) – entsprechend niederschlagen. Neben dem Markup von Dokumenten kann man mit Metadaten (*Schmidt* 2003) eine zusätzliche Informationsebene in den Analyseprozess integrieren. Eine Aufgabe der Computerlinguistik wird es sein, Verfahren zu entwickeln, diese expliziten Informationen über ein Dokument in intelligente Analysemethoden einzugliedern, wobei auch der automatische Aufbau von Metadaten – basierend auf dem Inhalt eines Dokuments – ein wichtiges Anwendungsszenario sprachverarbeitender Systeme ist, da Metadaten nur selten von Autoren in Dokumenten annotiert werden.

1.2.3 Aufbau von Korpora

Sollen HTML-Dokumente als Ressource innerhalb eines natürlichsprachlichen Systems eingesetzt werden, ist meist der Aufbau eines Korpus von Dokumenten unabdingbar (*Rehm* 2001). Es existieren zahlreiche frei verfügbare Werkzeuge, die für diesen Zweck benutzt und adaptiert werden können. In *Rehm* 2003a wird beispielsweise ein Korpus vorgestellt, das ca. 4 Mio. deutschsprachige Webseiten aus dem Bereich der universitären Webserver umfasst und insgesamt etwa 1,13 Milliarden Wortformen enthält.

In HTML-Dokumenten enthaltene Hyperlinks spannen einen Graph auf, der mit einem *Crawler* (häufig auch als *Spider* oder *Robot* und gelegentlich als *Agent* bezeichnet) traversiert werden kann (*Chakrabarti* 2003). Der Crawler transferiert das vom Benutzer anzugebende Startdokument per HTTP auf den lokalen Rechner, woraufhin alle Hyperlinks identifiziert und in eine Agenda eingetragen werden, die daraufhin sukzessive verarbeitet wird, um HTML-Dokumente sammeln zu können. Hierbei können bzw. müssen diverse Beschränkungen eingesetzt werden, damit dieses rekursive Verfahren terminiert, z. B. eine Obergrenze bzgl. der Dokumentanzahl oder eine Fokussierung auf die Dokumente einer be-

stimmten Domäne (beispielsweise nur `*.uni-giessen.de`). In den Prozess der Datensammlung können auch computerlinguistische Werkzeuge einbezogen werden, um spezielle Filterungsprozesse zu aktivieren. Hierbei werden automatische Sprachenidentifizierer eingesetzt, um nur diejenigen Dokumente in ein Korpus aufzunehmen, die in einer bestimmten Sprache verfasst worden sind (vgl. etwa *Cowie et al.* 1998 für Türkisch, Arabisch und Russisch, *Rehm* 2002 für Deutsch und *Jones und Ghani* 2000 für Tagalog). Nach der Datensammlung werden die Dokumente häufig indexiert, damit inhaltsorientierte Suchanfragen ermöglicht werden (vgl. Unterkapitel **??**, *Volltextsuche und Text Mining*).

Da verschiedene Versionen von HTML existieren und Dokumente häufig fehlerhaft annotiert sind, empfiehlt es sich, die Dateien vor der konkreten Verarbeitung zu normalisieren. Hierzu bietet sich eine Konvertierung arbiträrer HTML-Strukturen nach XHTML an, die innerhalb eines Tests mit 10 000 zufällig ausgewählten Dokumenten mit einer Präzision von ca. 98,7% durchgeführt werden konnte (*Rehm* 2003a). Da XHTML-Dateien zugleich XML-Instanzen sind, besteht der große Vorteil dieses Ansatzes darin, nachfolgenden Analysemodulen wohlgeformte (aber nicht notwendigerweise valide) XML-Instanzen zur Verfügung stellen zu können, was wiederum den Einsatz beliebiger XML-Werkzeuge ermöglicht (vgl. Unterkapitel **??**, *Texttechnologie* sowie *Myllymaki* 2001, *Rehm* 2003a, *Rehm* 2003b). Ein zweiter Vorteil betrifft die parallele Annotation (*Witt* 2003), da durch zusätzlich deklarierte *Namespaces* neben HTML weitere Annotationsebenen benutzt werden können, um z. B. Satzgrenzen oder Wortarteninformationen explizit mit Hilfe von XML-Elementen und -Attributen zu markieren.

1.2.4 Annotation linguistischer Informationen

Um sprachtechnologische Methoden in das Web integrieren zu können, wurde verschiedentlich vorgeschlagen, den Autoren von Webdokumenten Hilfsmittel, z. B. in Form spezieller Editoren, zur Verfügung zu stellen, damit eigene Dokumente mit syntaktischen, semantischen und pragmatischen Informationen annotiert und somit von Suchmaschinen detaillierter analysiert werden können. *Nagao und Hasida* 1998 stellen die *Global Document Annotation* (GDA) mit dem Anwendungsszenario des automatischen Textzusammenfassens vor (vgl. speziell hierzu *Rehm* 1999 sowie Unterkapitel **??**, *Textzusammenfassung*); Methoden des Textzusammenfassens operieren dabei auf den Informationen, die von den GDA-Elementen expliziert werden. *Dorai und Yacoob* 2002 skizzieren mit *Embedded Grammar Tags* (EGT) ein Verfahren, mit dem Autoren die potentiell antizipierbaren Suchanfragen an eine Webseite, z. B. bzgl. eines Dokuments mit Wetterinformationen, mit Hilfe eingebetteter XML-Elemente annotieren können:

```
<robotgram-in query="* [time] [does|did] [is] the sun [will|would]
(rise|rises|rose) * at College Park">06:19am</robotgram-in>.
```

Derartige manuelle Annotationen können aus den verschiedensten Gründen nicht großflächig erwartet werden, weshalb Verfahren zur robusten und insbesondere maschinellen Annotation linguistischer Informationen in Webdokumenten im Mittelpunkt der Web-orientierten Sprachtechnologie stehen.

Die von *Ide und Véronis* 1994 vorgeschlagenen Ebenen der manuellen Auszeichnung textueller Informationen (vgl. Unterkapitel **??**, *Texttechnologie*) kann man dabei unmittelbar auf die automatische Annotation übertragen, um z. B. die im Rahmen der GDA vorgeschlagene *manuelle* Markierung von Phrasenstrukturen mit Tagging- bzw. Parsing-Verfahren (siehe Unterkapitel **??**, *Syntax und Parsing*) zu *automatisieren.* Dabei ergibt sich eine Vielzahl interessanter Fragestellungen, die etwa die Anreicherung multilingualer HTML-Dokumente mit linguistischen Informationen betreffen, die Entwicklung robuster Systeme, die Merkmale für konzeptionelle Mündlichkeit (Smileys, spezifische Abkürzungen etc., *Haase et al.* 1997, *Rehm* 2003c) verarbeiten können, oder die Konzeptionierung neuartiger Anwendungen, z. B. Verfahren zur Filterung von HTML-Dokumenten nach ihren jeweiligen Hypertextsorten (*Rehm* 2003a, *Rehm* 2003b).

1.2.5 Szenarien der computerlinguistischen Nutzung des WWW

Im Kontext des World Wide Web existieren die verschiedensten computerlinguistischen und texttechnologischen Anwendungsszenarien. Im Kapitel „The Future of Web Mining“ seiner Web-Mining-Einführung führt *Chakrabarti* 2003 an, dass in Zukunft insbesondere Methoden der Sprachverarbeitung Einzug in dieses Gebiet finden werden, z. B. die Nutzung von Wortnetzen (vgl. Unterkapitel **??**, *Lexikalisch-semantische Wortnetze*), die Integration leistungsfähiger Part-of-Speech-Tagger oder der Einsatz spezialisierter Ontologien innerhalb domänenspezifischer Suchmaschinen. Zentrale, wenngleich nur mittel- bis langfristig realisierbare Bereiche sind weiterhin Parsing (vgl. Unterkapitel **??**, *Syntax und Parsing*) und die Akquisition und Repräsentation von Wissen basierend auf den Inhalten semistrukturierter Webseiten, wobei maschinelle Lernverfahren eine wesentliche Rolle spielen (*Soderland* 1997, *Chakrabarti* 2003).

Im Folgenden werden abschließend drei Anwendungen vorgestellt, die exemplarisch die Breite der Arbeiten mit dem World Wide Web charakterisieren sollen. *Heyer et al.* 2001 beschreiben das Projekt *Deutscher Wortschatz* (`http://www.wortschatz.uni-leipzig.de`), in dem u. a. HTML-Dokumente als Quellen zur Erstellung verschiedener monolingualer Korpora eingesetzt werden. Zahlreiche Filterkomponenten werden benutzt, um z. B. Eigennamen zu erkennen oder Kollokationen aufzubauen, die auch über eine Web-Schnittstelle visualisiert werden können. *Kunze und Rösner* 2003 stellen die Document Suite XDoc (*XML-basiertes Document Processing*) vor, die Experten den Zugriff auf Fachinformationen erleichtern soll, die u. a. aus Webseiten gewonnen werden. XDoc setzt auf verschiedenen Ebenen XML-basierte Formate ein, z. B. zur Repräsentation von Kasusrahmen oder zur Darstellung semantischer Analyseergebnisse des Chart-Parsers, die als XML Topic Map exportiert werden können (vgl. Unterkapitel **??**, *Texttechnologie*). XDoc nutzt auch in Webseiten enthaltene Strukturinformationen, so akzeptiert der Parser in Überschriften lediglich Nominalphrasen, wohingegen in Paragraphen mit Fließtext ganze Sätze erwartet werden. Das dritte Beispiel umfasst einen sehr speziellen Anwendungsbereich der Verarbeitung von Webseiten: die Analyse der von HTML-Tags aufgespannten Elementbäume, die potentiell Auskunft über die Struktur einer Seite geben

können, wobei jedoch zu beachten ist, dass in HTML logisch-strukturelle und präsentationsorientierte Auszeichnungsebenen vermischt werden, weshalb solche Analysen nur mit aufwendigen Verfahren sinnvolle Ergebnisse liefern können (vgl. u. a. *DiPasquo* 1998, *Chan und Yu* 1999, *Carchiolo et al.* 2000). Derartige Algorithmen werden insbesondere für die Informationsextraktion (vgl. Unterkapitel **??**, *Informationsextraktion*) eingesetzt, um spezifische Informationseinheiten in Webseiten lokalisieren zu können, um ein Dokument in logische Strukturmodule zu partitionieren oder zur Konvertierung von HTML-Dokumenten in andere Auszeichnungssprachen, z. B. WML (Wireless Markup Language, eingesetzt auf WAP-Endgeräten). Mittlerweile werden aus texttechnologischer Sicht zur Realisierung dieser Algorithmen häufig DOM-Prozessoren (*Document Object Model*, *Hors et al.* 2000, siehe auch Unterkapitel **??**, *Texttechnologie*) eingesetzt, vgl. *Chakrabarti* 2003 und *Rehm* 2003a.

1.3 Das Semantic Web

Tim Berners-Lee, Entwickler der Basistechnologien des WWW, verfolgt seit etwa 1998 die Vision des *Semantic Web*, das insbesondere auf explizit annotierten XML-Dokumenten sowie umfangreichen Metadatenbeschreibungen nach dem RDF- bzw. RDFS-Standard (vgl. Unterkapitel **??**, *Texttechnologie*) basieren soll: „a web of data that can be processed directly or indirectly by machines [...] We will solve large analytical problems by turning computer power loose on the hard data of the Semantic Web“ (*Berners-Lee* 1999, S. 177 ff.). Der aktuelle Stand des WWW manifestiert sich jedoch in Form äußerst unstrukturierter HTML-Dokumente (die Webserver *aller* deutschen Universitäten bieten etwa 4 Millionen HTML-Dokumente, aber nur ca. 25 000 XML-Dokumente an, *Rehm* 2003a). Der Weg zur immer umfangreicheren Strukturierung und damit auch zur effizienteren und robusteren maschinellen Auswertung von Dokumenten wird derzeit durch den Erfolg von XML und die Entwicklung flankierender Standards geebnet. Die großflächige Implementierung dieser Technologien – die Schaffung des Semantic Web – wird die aktuelle und zukünftige computerlinguistische Forschung immer stärker beeinflussen und zu neuen und innovativen Methoden und Anwendungsgebieten führen. Der Bedarf an computerlinguistischen Anwendungen ist, bewirkt durch den Erfolg des World Wide Web, größer denn je. Man kann sogar behaupten, dass nur sprachtechnologische Methoden in der Lage sein werden, die allgegenwärtige Informationsflut für den Endbenutzer in sinnvoller Weise eindämmen zu können.

1.4 Literaturhinweise

Auf den Seiten des W3C (`http://www.w3.org`) wird die Idee des Semantic Web diskutiert, Projekte und relevante Software werden unter `http://www.semanticweb.org` gebündelt aufgeführt. *Chakrabarti* 2003 enthält detaillierte Informationen zum Einsatz von Crawlern. Zahlreiche Beispiele für Anwendungen computerlinguistischer und texttechnologischer Verfahren auf das World Wide Web enthalten die Beiträge in *Lobin* 1999 und *Mehler und Lobin* 2003.

Literatur

Bergman, M. K. (2000): „The Deep Web: Surfacing Hidden Value“. White Paper, BrightPlanet.com LLC.

Berners-Lee, T. (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, T., Cailliau, R., Groff, J.-F. und Pollermann, B. (1992): „World-Wide Web: The Information Universe“. In: *Electronic Networking: Research, Applications and Policy*, **1**(2), 52–58.

Bray, T., Paoli, J., Sperberg-McQueen, C. M. und Maler, E. (2000): „Extensible Markup Language (XML) 1.0 (Second Edition)“. Technische Spezifikation, World Wide Web Consortium. Online verfügbar: `http://www.w3.org/TR/2000/REC-xml-20001006`.

Carchiolo, V., Longheu, A. und Malgeri, M. (2000): „Extracting Logical Schema from the Web“. In: Tan, A.-H. und Yu, P. S., Hrsg. , *Proceedings of the International Workshop on Text and Web Mining*, S. 64–71. Melbourne.

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

Chan, M. und Yu, G. (1999): „Extracting Web Design Knowledge: The Web DeCompiler“. In: *IEEE International Conference on Multimedia Computing and Systems (ICMCS 1999)*, Bd. 2, S. 547–552. IEEE Computer Society, Florence.

Cowie, J., Ludovik, E. und Zacharski, R. (1998): „An Autonomous, Web-based, Multilingual Corpus Collection Tool“. In: *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*, S. 142–148. Moncton. Online verfügbar: `http://crl.nmsu.edu/~raz/langrec/nlpia.htm`.

DiPasquo, D. (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, G. K. und Yacoob, Y. (2002): „Embedded Grammar Tags: Advancing Natural Language Interaction on the Web“. In: *IEEE Intelligent Systems*, **17**(1), 48–53.

Haase, M., Huber, M., Krumeich, A. und Rehm, G. (1997): „Internetkommunikation und Sprachwandel“. In: Weingarten, R., Hrsg. , *Sprachwandel durch Computer*, S. 51–85. Opladen: Westdeutscher Verlag.

Heyer, G., Läuter, M., Quasthoff, U. und Wolff, C. (2001): „Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse“. In: Lobin, H., Hrsg. , *Sprach- und Texttechnologie in digitalen Medien – Proceedings der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung*, S. 71–83. Gesellschaft für linguistische Datenverarbeitung, Justus-Liebig-Universität Gießen.

Hors, A. L., Hégaret, P. L., Wood, L., Nicol, G., Robie, J., Champion, M. und Byrne, S. (2000): „Document Object Model (DOM) Level 2 Core Specification“. Technische Spezifikation, World Wide Web Consortium.

Ide, N. und Véronis, J. (1994): „MULTEXT: Multilingual Text Tools and Corpora“. In: *COLING 94 – The 15th International Conference on Computational Linguistics*, Bd. 1, S. 588–592. Association for Computational Linguistics, Kyoto, Japan.

Jones, R. und Ghani, R. (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/`.

Kunze, M. und Rösner, D. (2003): „XDOC – XML-basierte Werkzeuge für die Extraktion und Repräsentation von Informationen“. In: *Mehler und Lobin* (2003). Erscheint.

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

Lobin, H. und Stührenberg, M. (2003): „XML-strukturierte Learning Objects“. In: Cyrus, L., Feddes, H., Schumacher, F. und Steiner, P., Hrsg. , *Sprache zwischen Theorie und Technologie – Festschrift für Wolf Papproté zum 60. Geburtstag*, Sprachwissenschaft, S. 185–198. Wiesbaden: Deutscher Universitäts-Verlag.

Lubell, J. (1999): „Structured Markup on the Web“. In: *Markup Languages*, **1**(3), 7–22.

Mehler, A. und Lobin, H., Hrsg. (2003): *Automatische Textanalyse – Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*. Wiesbaden: Westdeutscher Verlag. Erscheint.

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

Nagao, K. und Hasida, K. (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*, Bd. 2, S. 917–921. Association for Computational Linguistics, Montreal, Quebec, Kanada. 2 Bände.

Rehm, G. (1999): „Automatische Textannotation – Ein SGML- und DSSSL-basierter Ansatz zur angewandten Textlinguistik“. In: *Lobin* (1999), S. 179–195.

Rehm, G. (2001): „`korpus.html` – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten“. In: Lobin, H., Hrsg. , *Sprach- und Texttechnologie in digitalen Medien – Proceedings der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung*, S. 93–103. Gesellschaft für linguistische Datenverarbeitung, Justus-Liebig-Universität Gießen.

Rehm, G. (2002): „Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage“. In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii: IEEE.

Rehm, G. (2003a): „Hypertextsorten-Klassifikation als Grundlage generischer Informationsextraktion“. In: *Mehler und Lobin* (2003). Erscheint.

Rehm, G. (2003b): „Ontologie-basierte Hypertextsorten-Klassifikation“. In: *Mehler und Lobin* (2003). Erscheint.

Rehm, G. (2003c): „Schriftliche Mündlichkeit in der Sprache des World Wide Web“. In: Ziegler, A. und Dürscheid, C., Hrsg. , *Kommunikationsform E-Mail*, S. 263–308. Tübingen: Stauffenburg.

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 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, Paulk J. Leach und Tim Berners-Lee. Online verfügbar: `http://www.ietf.org/rfc/`.

Schmidt, I. (2003): „Modellierung von Metadaten“. In: Lobin, H. und Lemnitzer, L., Hrsg. , *Texttechnologie – Anwendungen und Perspektiven*, S. 143–164. Tübingen: Stauffenburg. Erscheint.

Soderland, S. (1997): „Learning to Extract Text-Based Information from the World Wide Web“. In: Heckerman, D., Mannila, H. und Pregibon, D., Hrsg. , *Proceedings of the 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-97)*, S. 251–254. Newport Beach: AAAI Press.

Witt, A. (2003): „Linguistische Informationsmodellierung mit XML“. In: *Mehler und Lobin* (2003). Erscheint.