Computerlinguistik, Markup-Sprachen, und das World Wide Web

Georg Rehm

Justus-Liebig-Universität Giessen
Arbeitsbereich Angewandte Sprachwissenschaft und Computerlinguistik
Otto-Behaghel-Strasse 10 D
35394 Giessen

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

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

28. September 2000

1 Einleitung

Im Januar 2000 haben etwa 6,5 Millionen *World Wide Web*-Server weltweit mehr als 1 Milliarde Dokumente angeboten, aktuellere Hochrechnungen sprechen schon von mehr als 2,4 Milliarden Dokumenten.[1] Da niemand diese Massen an tagesaktuellen Informationen, Nachrichten, Essays, Artikeln und Geschichten in sinnvoller Weise strukturieren oder gar rezipieren kann, befindet sich die Anwendung computerlinguistischer Methoden (etwa zur Informationsextraktion, zum Aufbau von Thesauri oder zur Verbesserung von Suchmaschinentechnologien) momentan im Mittelpunkt der informationstechnologischen Forschung. Des weiteren werden verschiedene Unzulänglichkeiten der *Hypertext Markup Language* (HTML) hinsichtlich einer flexiblen Strukturierung beliebiger Inhalte als *die* Ursache für die Schwierigkeiten beim Aufbau effizienter Suchmaschinen und intuitiver Methoden zur Datenexploration und -navigation angesehen. Aus diesem Grund hat das *World Wide Web Consortium*[2] die *Extensible Markup Language* (XML) entwickelt. Da mit Hilfe von XML geschaffene Definitionen von *Markup*-Sprachen regelbasiert und mit kontextfreien Grammatiken zur Beschreibung natürlicher Sprachen vergleichbar sind, kommt unmittelbar die computerlinguistische Arbeit ins Spiel, denn Aufbau und Pflege von *Dokumenttyp-Definitionen* sind aufwändige Prozesse, die sowohl generelles semantisches und textlinguistisches Wissen als auch die Kenntnis der zugrundeliegenden Textsorten, Auszeichnungssprachen und auch Skript- und Verarbeitungssprachen erfordern.

2 Markup-Sprachen als Fundament des WWW

Dieser Abschnitt skizziert zunächst die *Lingua Franca* des WWW, die Hypertext Markup Language sowie die Extensible Markup Language. Anschließend wird mit dem *Resource Description Framework* exemplarisch ein Formalismus vorgestellt, der zu denjenigen Standards gehört, die XML flankieren, d. h. in XML selbst definiert sind und funktionale Schwerpunkte setzen.

2.1 HTML: Hypertext Markup Language

HTML (*Raggett et al.* 1999) gestattet die Anreicherung einer ASCII-Textdatei mit Auszeichnungs-*Elementen*. So markiert beispielsweise das Element `<P>` (*paragraph*) den Beginn eines Absatzes, `</TABLE>` das Ende einer Tabelle, und Text, der von den Elementen `<H1>` und `</H1>` (*headline*) eingeschlossen wird, stellt eine Überschrift erster Stufe dar. Weitere Elemente erlauben die Auszeichnung zusätzlicher Ebenen von Überschriften, verschiedener Arten von Listen und vor allem die Integration von *Hyperlinks*. Diese Elemente sind nicht in beliebiger Form kombinierbar, sondern es existiert eine regelbasierte, formale Definition, die die Namen und das Zusammenspiel der Elemente und ihrer Attribute in Form einer Grammatik spezifiziert. Diese *Dokumenttyp-Definition* (DTD, *Document*

[1] Siehe `http://www.inktomi.com/webmap/` sowie `http://censorware.org/web_size/`.
[2] Das W3C verabschiedet neue Standards im Bereich der Web-Technologien.

Type Definition) für HTML wurde mit Hilfe der *Standard Generalized Markup Language* (SGML, *ISO8879* 1986) definiert.

2.2 XML: Extensible Markup Language

Durch den Erfolg des World Wide Web und die explosionsartig wachsende Menge an Dokumenten war bereits Anfang 1996 absehbar, dass die Grenzen von HTML erreicht waren. Diese Grenzen beziehen sich vornehmlich auf die mangelnde Erweiterbarkeit: Da es sich bei HTML um eine Anwendung von SGML handelt, sind Namen, Anzahl und Möglichkeiten der Kombinationen von Auszeichnungselementen vordefiniert und somit nicht vom Benutzer veränderbar.[3] Gerade dieser HTML inhärente Mangel an Möglichkeiten zur *expliziten Strukturierung beliebiger Informationen* war für das W3C der Anlaß, die Extensible Markup Language (*Bray et al.* 1998) zu spezifizieren.[4] XML gestattet diese Explizierung arbiträrer Informationen, so dass eine effiziente und sinnvolle automatische Verarbeitung von Dokumenten, etwa zu Recherchezwecken oder zur Anpassung eines Dokuments an die Wünsche der Benutzer durch eine *on the fly* durchgeführte Sortierung oder Filterung, prinzipiell gewährleistet ist.

Bereits 1994 wurde vorgeschlagen, die formale Mächtigkeit von SGML auf das WWW zu übertragen, um die Eingeschränktheit von HTML zu überwinden (*Sperberg-McQueen und Goldstein* 1994). Der Erfolg von HTML hatte gezeigt, dass eine SGML-basierte Sprache für einen Einsatz in verteilten Netzwerken durchaus geeignet ist. Der unmittelbaren Übertragung von SGML in das WWW stand jedoch dessen große Komplexität gegenüber (*Goldfarb* 1999). Mit der Entwicklung von XML wurde durch eine konsequente Kürzung der Definition von SGML ein Formalismus geschaffen, der die für den Einsatz im Internet wichtigsten definitorischen Eigenschaften von SGML enthält, ohne jedoch dessen Ausdrucksfähigkeit in prinzipieller Weise einzuschränken (*Rehm und Lobin* 2000). Da eine Einführung in die Syntax von XML den Umfang dieses Beitrags sprengte, soll Abbildung 1 den Vorteil von XML gegenüber HTML – die Explizierung von Struktur – veranschaulichen.[5]

Die strukturelle Freiheit von XML birgt jedoch auch Gefahren: Da die Namen von XML-Elementen frei definierbar sind, wird eine Vielzahl von Auszeichnungssprachen für identische oder sehr ähnliche Einsatzgebiete bzw. Dokumenttypen entstehen. So könnten Autoren von Dokumenttyp-Definitionen, die die logische Struktur eines Adressbuches beschreiben, als Element für die Strassenadresse eben nicht `<strassenadresse>` sondern beispielsweise `<stradr>` oder `<adr>` benutzen. War das zentrale Problem von HTML ein *Mangel* an expliziter Struktur, wird die Zukunft also möglicherweise einen *Überschuss* an Struktur

[3] In HTML können etwa Literaturangaben nicht mit expliziten Elementen wie `<author>`, oder `<publisher>` annotiert werden, da diese nicht in der HTML-DTD vorgesehen sind.

[4] XML ist eine Teilmenge der Standard Generalized Markup Language. SGML geht auf die Idee des *generischen Markup* (*Gennusa* 1999) zurück: SGML dient der plattformunabhängigen Definition beliebiger Auszeichnungssprachen, wobei von der visuellen Präsentation textueller Daten zugunsten einer Explizierung von Struktur und Bedeutung abstrahiert wird.

[5] Nach *Witt* 1999, der SGML aus (computer)linguistischer Perspektive betrachtet.

```
<?XML version=‘‘1.0’’?>
<S>
    <NP KASUS=‘‘NOM’’>
        <DET>Der</DET>
        <N>Mann</N>
    </NP>
    <VP>
        <V PERSON=‘‘3’’ TEMPUS=‘‘PRAES’’>tritt</V>
        <NP KASUS=‘‘AKK’’>
            <DET>den</DET>
            <N>Ball</N>
        </NP>
    </VP>
</S>
```

Abbildung 1: Ein XML-Dokument am Beispiel von „Der Mann tritt den Ball.“

hevorbringen. Diesem Problem kann man mit unterschiedlichen Methoden begegnen. So wäre beispielsweise denkbar, dass ähnlich dem *Open Source* Ansatz in der Software-Entwicklung eine globale Gemeinschaft entsteht, die sich der freien Entwicklung und Pflege von Markup-Sprachen widmet (*Rehm und Lobin* 2000).

Eine weitere Möglichkeit bestünde darin, ein sprachverarbeitendes System zu entwickeln, das mit einem sehr speziellen Fragment natürlicher Sprachen umgehen kann: den Namen von XML-Elementen einer oder mehrerer spezialisierter Domänen wie z. B. Adressbüchern. Ein solches System könnte die morpholexikalischen Ähnlichkeiten von Elementnamen wie `<strassenadresse>`, `<stradr>` und `<adr>` automatisch erkennen und die jeweiligen durch die unterschiedlichen Dokument-Grammatiken definierten Inhalte (weitere Elemente oder konkrete textuelle Daten) einander zuordnen oder bei der manuellen Zuordnung korrespondierender Elemente unterstützend wirken. Dies könnte man mit einfachen *Pattern-Matching* Methoden, einem Lexikon und einer Kompositaanalyse realisieren. Ab einer bestimmten Detailstufe müsste ein derartiges System auch über eine gewisse Menge Weltwissen verfügen, um die Teil-Ganzes-Relationen von Elementen wie etwa `<vorwahl>` und `<durchwahl>` in Bezug auf `<telefonnummer>` zu erkennen. Generellere Hyperonym/Hyponym-Relationen zwischen Begriffen sind hierbei durch Konzepthierarchien wie beispielsweise (Euro)WordNet auflösbar.

2.3 Metadaten: Das Resource Description Framework

Der vom W3C vorgeschlagene Formalismus zur Annotation von Metadaten ist das XML-basierte Resource Description Framework (RDF, *Lassila und Swick* 1999). Metadaten sind beispielsweise der Name des Autors, das Datum der letzten Änderung oder Schlagwörter.

RDF gestattet, ähnlich wie XML, die Definition von Schemata, mit deren Hilfe dann wiederum Dokumente annotiert werden können. Hierbei ergeben sich verschiedene Fragestellungen: Welche Vokabulare werden zur Definition einge-

setzt? Wie detailliert werden die Metadaten strukturiert, welche Metadaten sollen annotierbar sein? Zu dieser Problematik gibt es in verschiedenen Fachrichtungen (Bibliothekswesen, Architektur, Kunst etc.) Bemühungen zur Schaffung von Standards (*Baca* 1998). Im WWW scheint sich die noch in der Entwicklung befindliche Initiative *Dublin Core* (`http://purl.org/dc/`) durchzusetzen, die einen erweiterbaren Kern zur Metadaten-Auszeichnung hervorbringen soll.

Das im vergangenen Abschnitt knapp umrissene System zur Realisierung einer bidirektionalen Abbildung verschiedener DTDs, die *unterschiedliche* Element-Namen definieren, jedoch eine *identische* Domäne beschreiben, könnte man mit bereits verfügbaren RDF-Schemata verbessern: Das *Open Directory Project* (`http://dmoz.org`) erstellt ein möglichst umfassendes Verzeichnis des WWW, das ähnlich wie *Yahoo* aufgebaut ist. Die Themenbereiche werden in Form eines mehrere hunderttausend Einträge umfassenden RDF-Schemas verwaltet, das prinzipiell einer Konzepthierarchie entspricht, die – ebenso wie WordNet – zur Entdeckung von Hyperonym- und Hyponym-Relationen eingesetzt werden kann.

3 Computerlinguistik und das World Wide Web

Man kann verschiedene Bereiche identifizieren, in denen die Computerlinguistik einen nicht unbeträchtlichen Einfluß auf die zukünftige Entwicklung des WWW und intelligentere und intuitivere Arten der Interaktion mit diesem Medium haben wird. Zunächst skizzieren wir jedoch, in welcher Form (computer)linguistische Forschungsvorhaben mit SGML und XML arbeiten.

3.1 Der computerlinguistische Einsatz von SGML/XML

Die Computerlinguistik setzt SGML und XML in verschiedenen Bereichen ein. Das prominenteste Beispiel ist die Arbeit der *Text Encoding Initiative*, TEI. Diese Gruppe von mehr als 100 Wissenschaftlern aus der Literatur- und Sprachwissenschaft, der Informatik und Computerlinguistik hat über mehrere Jahre hinweg modular aufgebaute SGML Dokumenttyp-Definitionen zur Auszeichnung von Sprache (für so unterschiedliche Textsorten wie Gedichte, Dramen, historische Materialien, Wörterbücher etc.) erarbeitet und als zweibändiges, Kompendium veröffentlicht (*Sperberg-McQueen und Burnard* 1994). Die TEI-DTDs werden – oftmals in erweiterter Form – für die Annotation großer Korpora, aber etwa auch für die Transkription gesprochener Sprache eingesetzt.

Weitere Anwendungsgebiete von SGML sind die automatische Erkennung von Dokumentstrukturen (*Huck et al.* 1998), Sammlungen von SGML- und XML-fähigen UNIX-Werkzeugen zur automatischen Korpus-Annotation und -Auswertung (*McKelvie et al.* 1997) oder der Einsatz von SGML zur manuellen Markierung rhetorischer Strukturen, um eine flexible und benutzer-angepasste Präsentation in einer Hypermedia-Umgebung zu gewährleisten (*Lobin* 1999a).

Im Rahmen des MULTEXT-Projekts (*Ide und Véronis* 1994) werden vier Ebenen der Auszeichnung textueller Daten unterschieden: Auf der ersten Ebene kann man dokumentweite Auszeichnungen vornehmen, Dokumente bibliogra-

phisch erfassen, verwendete Zeichensätze aufführen etc. Die zweite Ebene beinhaltet eine Auszeichnung textueller Einheiten wie Band, Kapitel, Abschnitt oder Fußnote. Auf der dritten Ebene werden Strukturen innerhalb von Abschnitten explizit markiert: Sätze, Wörter, Abkürzungen oder Eigennamen. Die vierte, detailreichste Ebene umfasst schließlich die Markierung syntaktischer Kategorien oder morphologischer Einheiten.

Die aktuelle Sichtweise bzgl. des Einsatzes von SGML/XML in traditionellen computerlinguistischen Anwendungsgebieten konzentriert sich, wie o. a. Beispiele deutlich machen, auf den Bereich der „textuellen Datenbanken“ (*Lobin* 1999a, *Hockey* 1998) und eine Auswertung und Aufbereitung dieser linguistischen Datenbestände auf den von *Ide und Véronis* 1994 angenommenen vier Ebenen (*Nerbonne* 1998, *Lobin* 1999b). Dabei geht es um die manuelle, semi-automatische oder automatische Annotation von Texten nach unterschiedlich detaillierten linguistischen Kriterien, eine Auswertung des annotierten Materials und auch um die Anwendung dieser computerlinguistischen und texttechnologischen Methoden auf konkrete Anwendungsszenarien (*Möhr und Schmidt* 1999).

3.2 Computerlinguistik, XML und das WWW

Die angesprochenen Verfahren zur Anreicherung und Verarbeitung von SGML- bzw. XML-Dokumenten werden aufgrund der identischen Herkunft der zugrundeliegenden Auszeichnungssprachen auch im WWW eingesetzt: Annotierte Korpora können beispielsweise in statische HTML-Dokumente konvertiert werden, und auch dynamische Konvertierungen sind in vielfältiger Hinsicht realisierbar (*Lubell* 1999). Verlage bedienen sich dieser Methoden, um bestehende Substanzen im Zuge des *Electronic* bzw. *Cross Media Publishing* für das WWW aufzubereiten (*Möhr und Schmidt* 1999).

Die Anwendung computerlinguistischer Methoden beschränkt sich hierbei auf den automatischen Aufbau von Annotationen[6] bzw. einen intelligenten und eher texttechnologisch motivierten Umgang mit bestehenden SGML/XML-Dokumenten, um diese etwa in einer WWW-basierten adaptiven Hypermedia-Umgebung abhängig vom Wissensstand des Benutzers präsentieren zu können.

Bislang finden sich in der Literatur nur sporadische Ansätze für einen unmittelbaren Einsatz der Kombination von Auszeichnungssprachen und Computerlinguistik in Bezug auf das WWW. *Nagao und Hasida* 1998 schlagen die *Global Document Annotation* vor: Da HTML sich lediglich auf das Layout eines Dokuments beziehe, müsse man die Autoren von HTML-Dokumenten dazu bewegen, ihre Texte mittels eines speziellen Editors mit den GDA-Elementen anzureichern, die vornehmlich syntaktische, semantische und rhetorische Informationen explizieren.[7] Doch wie eingangs erwähnt wurde, besteht das WWW mittlerweile aus mehr als 2 Milliarden Dokumenten und kaum ein Autor wird einen weiteren Editor in den Produktionsprozess seiner Dokumente einbetten und sich das zur

[6] Ein *Part of Speech*-Tagger kann einzelnen Wörtern ihre syntaktischen Kategorien zuweisen und diese mit anderen Informationen in XML-Elementen hinterlegen.

[7] Das Anwendungsszenario ist hier das automatische Textzusammenfassen basierend auf den Informationen, die von den GDA-Elementen expliziert werden (vgl. *Rehm* 1999)

Bedienung notwendige linguistische Wissen aneignen, so dass das avisierte Ziel, die „globale und interkulturelle Kommunikation zu revolutionieren“ (*Nagao und Hasida* 1998), kaum als realistisch eingestuft werden kann.

Daher sollte sich die computerlinguistische Forschung vor allem mit der Frage beschäftigen, wie man bereits existierende HTML- (und in Zukunft auch XML-) Dokumente in sowohl möglichst umfassender als auch robuster Weise analysieren und automatisch annotieren kann. Die von *Ide und Véronis* 1994 vorgeschlagenen Ebenen der manuellen Auszeichnung textueller Informationen mit SGML-basierten Auszeichnungssprachen kann man dabei unmittelbar auf die automatische Annotation übertragen, um die im Rahmen des GDA-Projekts vorgeschlagene *manuelle* Markierung von Phrasenstrukturen mit Tagging- bzw. Parsing-Verfahren zu *automatisieren*. Dabei ergibt sich eine Vielzahl von herausfordernden Fragestellungen, die an dieser Stelle nur angerissen werden können: Wie werden linguistische Informationen in bestehende HTML-Dokumente integriert? Wie baut man konsistente Korpora von HTML-Dokumenten bzw. ganzen Web-Servern auf, und wie kann man Revisionen von Dokumenten automatisch in das Korpus integrieren (*Walker* 1999)? Wie kann man sprachverarbeitende Systeme derart robust gestalten, dass sie mit den vielen (para)linguistischen Elementen heutiger HTML-Dokumente und der oftmals vorherrschenden konzeptionellen Mündlichkeit (*Haase et al.* 1997) umgehen können?

Man kann sprachverarbeitende Systeme aber auch mit Informationen über HTML und XML 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 als Überschrift markierter Text vermutlich kein vollständiger Satz sondern lediglich eine NP sein dürfte, um daraufhin spezielle Regeln zur Analyse elliptischer Ausdrücke anzuwenden. Neben der textuellen Information ist also diese metasprachliche Ebene für die Unterstützung des automatischen Analyseprozesses immens wichtig und sollte sich in zukünftigen Systemen und Theorien entsprechend niederschlagen. Neben dem Markup von Dokumenten kann man eine weitere Ebene von metasprachlichen Informationen in den Analyseprozess integrieren: Metadaten. Metadaten werden in Zukunft immer seltener mit dem sehr unspezifischen HTML-Element `<meta>` ausgezeichnet werden; stattdessen wird das Resource Description Framework Verwendung finden. Sobald HTML- oder XML-Dokumente mit Metadaten ausgezeichnet werden, sollte es Aufgabe der Computerlinguistik sein, Verfahren zu entwickeln, diese expliziten Informationen über ein Dokument in intelligente Analysemethoden einzugliedern. Auch der automatische Aufbau von Metadaten basierend auf dem Inhalt eines Dokuments ist ein denkbares Anwendungsszenario zukünftiger sprachverarbeitender Systeme, deren Domäne die Analyse von HTML- und XML-Dokumenten sein wird.

4 Computerlinguistik und das Semantic Web

Tim Berners-Lee, Entwickler der Basistechnologien des WWW, verfolgt mittlerweile eine Vision namens *Semantic Web*: „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. Der Weg zur immer umfangreicheren Strukturierung, 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 könnte sogar behaupten, dass nur die Computerlinguistik und die sprachverarbeitende KI in der Lage sein werden, die allgegenwärtige Informationsflut für den Endbenutzer in sinnvoller Weise einzudämmen. Um die Vision vom Semantic Web zu realisieren, ist es eine *der* Herausforderungen der Computerlinguistik, sich den neuen Fakten und neuen Standards zu stellen und, hierauf basierend, neue Konzepte und Methoden zu entwickeln.

Literatur

Baca, M., Hrsg. (1998): *Introduction to Metadata – Pathways to Digital Information.* Getty Information Institute.

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.

Bray, T., Paoli, J. und Sperberg-McQueen, C. M. (1998): „Extensible Markup Language (XML) 1.0“. Technische Spezifikation, World Wide Web Consortium. Online verfügbar: `http://www.w3.org/TR/1998/REC-xml-19980210`.

Gennusa, P. L. (1999): „Evolution and Use of Generic Markup Languages“. In: Möhr, W. und Schmidt, I., Hrsg. , *SGML and XML – Anwendungen und Perspektiven*, S. 27–50. Berlin, Heidelberg, New York etc.: Springer.

Goldfarb, C. (1999): „Future Directions in SGML/XML“. In: Möhr, W. und Schmidt, I., Hrsg. , *SGML und XML – Anwendungen und Perspektiven*, S. 3–25. Berlin, Heidelberg, New York etc.: Springer.

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.

Hockey, S. (1998): „Textual Databases“. In: Lawler, J. M. und Dry, H. A., Hrsg. , *Using Computers in Linguistics: A Practical Guide*, S. 101–137. New York: Routledge.

Huck, G., Fankhauser, P., Aberer, K. und Neuhold, E. (1998): „Jedi: Extracting and Synthesizing Information from the Web“. In: *Third IFCIS Conference on Cooperative Information Systems – CoopIS'98.* IEEE Computer Society Press. Online verfügbar: `ftp://ftp.darmstadt.gmd.de/pub/oasys/reports/P-98-11.pdf`.

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. 2 Bände.

ISO8879 (1986): „Information Processing – Text and Office Information Systems – Standard Generalized Markup Language“. Internationaler Standard, Genf, International Organization for Standardization.

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

Lobin, H. (1999a): „Intelligente Dokumente – Linguistische Repräsentation komplexer Inhalte für die hypermediale Wissensvermittlung“. In: Lobin, H., Hrsg. , *Text im digitalen Medium – Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering*, S. 155–178. Wiesbaden: Westdeutscher Verlag.

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

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

McKelvie, D., Brew, C. und Thompson, H. (1997): „Using SGML as a Basis for Data-Intensive NLP“. In: *Proceedings of ANLP 97.* Association for Computational Linguistics, Washington D. C. Online verfügbar: `http://www.ltg.hcrc.ed.ac.uk/~dmck/anlp97.ps`.

Möhr, W. und Schmidt, I., Hrsg. (1999): *SGML und XML – Anwendungen und Perspektiven.* Berlin, Heidelberg, New York etc.: Springer.

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.

Nerbonne, J., Hrsg. (1998): *Linguistic Databases.* Nr. 77, CSLI Lecture Notes. Cambridge: Cambridge University Press.

Raggett, D., Hors, A. L. und Jacbos, I. (1999): „HTML 4.01 Specification“. Technische Spezifikation, World Wide Web Consortium. Online verfügbar: `http://www.w3.org/TR/html401/`.

Rehm, G. (1999): „Automatische Textannotation – Ein SGML- und DSSSL-basierter Ansatz zur angewandten Textlinguistik“. In: Lobin, H., Hrsg. , *Text im digitalen Medium – Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering*, S. 179–195. Wiesbaden: Westdeutscher Verlag.

Rehm, G. und Lobin, H. (2000): „From Open Source to Open Information – Collaborative Methods in Creating XML-based Markup Languages“. In: *Proceedings of Electronic Publishing 2000.* International Federation for Information Processing and International Council for Computer Communication, Kaliningrad, Svetlogorsk.

Sperberg-McQueen, C. M. und Burnard, L., Hrsg. (1994): *Guidelines for Electronic Text Encoding and Interchange.* Chicago, Oxford: University of Chicago, University of Oxford. Version P3, 2 Bände.

Sperberg-McQueen, C. M. und Goldstein, R. 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`.

Walker, D. (1999): „Taking Snapshots of the Web with a TEI Camera“. In: *Computers and the Humanities*, **33**, 185–192.

Witt, A. (1999): „SGML und Linguistik“. In: Lobin, H., Hrsg. , *Text im digitalen Medium – Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering*, S. 121–154. Wiesbaden: Westdeutscher Verlag.

Index