Originalveröffentlichung in: Lobin, Henning (Hrsg.): Text im digitalen Medium. Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering. - Opladen (u.a.): Westdeutscher Verlag, 1999. S. 1-7.

Textdesign, Texttechnologie und Hypertext Engineering

Einleitung in den vorliegenden Band

Henning Lobin

1 Text im digitalen Medium

Die Nutzung von Texten im digitalen Medium hat zur Entstehung von neuen Themen für die Linguistik geführt, die durch die Schlagwörter *Textdesign*, *Texttechnologie* und *Hypertext Engineering* bezeichnet werden können. Alle drei Themenbereiche sind – wie der vorliegende Band zeigen möchte – eng miteinander verbunden. Verbindungslinien werden einerseits durch den Aspekt der inhaltlichen Kohärenz, andererseits durch den der strukturellen Textauszeichnung hergestellt.[1]

Hans-Jürgen Bucher untersucht in seinem Beitrag „Die Zeitung als Hypertext – Verstehensprobleme und Gestaltungsprinzipien für Online-Zeitungen" die Umsetzung des über Jahrhunderte gewachsenen Erscheinungsbildes gedruckter Zeitungen im *World Wide Web*. Er fokussiert dabei insbesondere, wie online-spezifische Verstehensprobleme, die bei der Textsorte ‚Zeitung' auftreten können, bewältigt werden, um daraus Elemente einer Hypertext-Rhetorik für Online-Zeitungen abzuleiten.

Auch *Angelika Storrer* stellt traditionelle Texte und Hypertexte einander entgegen, allerdings unter dem Aspekt von „Kohärenz in Text und Hypertext". Nach wertvollen begrifflichen Klarstellungen und einer Übersicht über den textlinguistischen Kohärenzbegriff untersucht sie die Frage, inwieweit es sinnvoll ist, auch in Hypertexten von Kohärenz zu sprechen und welche Kohärenzbildungshilfen eingesetzt werden. Exemplifiziert wird die Untersuchung am grammatischen Informationssystem GRAMMIS.

---

[1] Ein Teil der Beiträge ist aus Vorträgen hervorgegangen, die auf der Jahrestagung der Gesellschaft für Angewandte Linguistik 1997 in Bielefeld in einem Themenbereich mit dem Titel „Sprache – Das Multimedium" gehalten worden sind. Für Unterstützung bei der Fertigstellung dieses Bandes möchte ich Jan-Torsten Milde und Andreas Witt danken.

Jürgen Handke konzentriert seinen Beitrag auf die „Präsentation und Lernerunterstützung im wissenschaftlichen Lernsystem", und zwar anhand des Systems *Linguistics Interactive*. Er erläutert, mit welchen Mitteln komplizierte linguistische Prozesse auf der Grundlage der heutigen Multimedia-Technologie so umgesetzt werden können, daß damit ‚didaktischer Mehrwert' gegenüber traditionellen Lehrformen erzielt werden kann.

Die Hinwendung zu strukturellen Aspekten geschieht mit dem Beitrag „Strukturierungsmethoden für Hypermediadokumente und ihre Umsetzung" von *Gerhard Heyer* und *Christian Wolff*. Ihr Beitrag beginnt mit einer Bestandsaufnahme der Analyse- und Bewertungsverfahren für Hypermedia und der zur Realisierung zur Verfügung stehenden Werkzeuge und Standards. Den Umgang mit der diagnostisierten sehr unübersichtliche Situation in diesem Bereich erläutern sie anhand des *Multimedialen Physikalischen Praktikums*, einem elektronischen Buch.

Die abschließende Darstellung weiterführender Standardisierungen im Beitrag von Heyer und Wolff leitet bereits zu dem Überblicksbeitrag „SGML und Linguistik" von *Andreas Witt* über. Witt stellt die *Standard Generlized Markup Language* (SGML), einem Standard zur inhaltlich-strukturellen Textauszeichnung, in seinen Einzelheiten dar und erläutert auch seine oft mißverstandenen Bezüge zur *Hypertext Markup Language* (HTML) und zur *Extensible Markup Language* (XML). Witt demonstriert auch, wie diese Standards in der Korpuslinguistik genutzt werden können.

Henning Lobin behandelt in seinem Beitrag „Intelligente Dokumente – Linguistische Repräsentation komplexer Inhalte für die hypermediale Wissensvermittlung" die SGML-Kodierung inhaltlicher Textstrukturen zum Zwecke einer flexiblen Inhaltspräsentation in komplexen Wissensbereichen. Grundlage ist dabei die *Rhetorical Structure Theory* (RST), die in diesem Beitrag hypermedial interpretiert wird und so eine übergreifende Basis für Text- und Hypertext-Strukturen bildet.

Auch *Georg Rehm* befaßt sich im Rahmen der „Automatischen Textannotation" mit der Verbindung von SGML und RST. Mit seinem „SGML- und DSSSL-basierten Ansatz zur angewandten Textlinguistik" (Untertitel) ist es ihm möglich, aus realen Zeitungstexten automatisch RST-Strukturen abzuleiten und diese nach mehreren Verarbeitungsschritten für die Generierung von Zusammenfassungen oder hypertextualisierten Formen der Textpräsentation zu nutzen.

Der den Band abschließende Beitrag von *Jan-Torsten Milde*, „Effizientes Document Engineering sprachlicher Daten", behandelt in noch stärkerem Maße den Aspekt der

Systementwicklung. Milde stellt einen Ansatz vor, SGML-kodierte Texte über verbreitete Datenbank-Techniken verfügbar zu halten und sie programmgesteuert in verschiedene Ausgabeformate zu überführen. Er zeigt dies anhand von linguistisch strukturierten Textdaten.

2 Linguistik und Informationstechnologie

Die Übersicht im vorangegangenen Abschnitt hat gezeigt, daß neben dem Aspekt der (Hypertext-)Kohärenz die Textstrukturierung ein durchlaufendes Thema in den Beiträgen dieses Bandes darstellt. Während das erste Querschnittsthema als ein linguistisches Thema mittlerweile anerkannt ist (vgl. vor allem die Literaturhinweise im Beitrag von Storrer in diesem Band), so ist die Relevanz der Verbindung von Linguistik und Informationstechnologie, speziell in ihrer Ausprägung auf der Basis von SGML, bislang noch nicht so deutlich hervorgehoben worden. In diesem Abschnitt sollen deshalb zu dieser Verbindung einige Bemerkungen erfolgen.

Bei der Repräsentation von Texten für die verschiedenen Zwecke der Verarbeitung in digitalen Medien hat sich seit geraumer Zeit SGML als ein Ansatz in den Vordergrund geschoben, der im Gegensatz zur traditionellen Textverarbeitung auf einer strikten Trennung von Inhalt, Form und Struktur beruht (vgl. den Beitrag von Witt in diesem Band). Texte werden dabei mit Markierungen versehen, die ihre Struktur aus einer bestimmten Perspektive explizit zu machen erlauben, so daß die äußere Gestaltung je nach Zielmedium, Gestaltungsanspruch, Verarbeitungstechniken und Rezipienten als ein unabhängiges und wiederverwendbares Modul entwickelt werden kann. Eine derartige Trennung von Inhalt, Struktur und Form hat sich als so fruchtbar erwiesen, daß diese Technologie inzwischen nicht nur für Textdokumente verwendet wird, sondern beispielsweise – wie der Erfolg des *World Wide Web* (WWW) zeigt – auch für multimediale Hypertextdokumente oder sogar vollkommen 'untextuelle' Daten, etwa technische Spezifikationen. Der Vorteil der Verwendung von SGML bei textuellen Daten liegt in der erhöhten Verfügbarkeit dieser Daten. Entscheidend für die Leistungen von SGML ist, daß die Strukturinformation zu einem bestimmten Datentyp durch eine Grammatik beschrieben wird, und zwar durch eine gewöhnliche kontextfreie Grammatik. Diese Struktur-Grammatik, die *Document Type Definition* (DTD), bildet den wesentlichen Unterschied zu der rein listenförmigen Aneinanderreihung der Daten in herkömmlichen Textdokumenten oder der strikt tabellarischen in Datenbanken. Die

DTD erlaubt die Beschreibung weitaus komplexerer Datenstrukturen als die der Liste oder der Tabelle in allgemeiner Form. Der Hauptunterschied von SGML-Editoren, -Datenbanken, -Konvertern und anderen SGML-Softwaresystemen besteht deshalb darin, diese Strukturgrammatik für die Verarbeitungszwecke des Editierens, Datenverwaltens oder Konvertierens gewinnbringend nutzen zu können.

Die Erstellung von Strukturgrammatiken, DTDs, ist eine Aufgabe, die grob gesehen drei Schritte umfaßt:

1. Bereits vorhandene Daten sind nach strukturellen Gesetzmäßigkeiten zu analysieren,
2. die strukturellen Anforderungen an zukünftige Daten sind zu explorieren, und
3. die Struktur in Form der Grammatik ist zu spezifizieren und im Anwendungsszenario auszutesten und zu revidieren.

Obwohl für den gesamten Ablauf der Implementation von SGML mittlerweile einige Gesamtdarstellungen vorliegen (vor allem Travis/Waldt 1995 und Maler/El Andaloussi 1996), fehlt für die zentrale Aufgabe der DTD-Erstellung bis heute eine verläßliche Methodologie. Die in den letzten Jahren durchgeführten oder immer noch laufenden Projekte zur DTD-Normierung in bestimmten Bereichen (für literarische Text die 'Text Encoding Initiative', vgl. Sperberg-McQueen/Burnard 1994) haben gezeigt, daß die DTD-Erstellung von so entscheidender Bedeutung, daß für die Durchführung nicht nur modernes Projekt-Management zwingend erforderlich ist, sondern für die Maximierung der Verläßlichkeit der Resultate auch Ergebnisse aus dem Bereich des *Software Engineering* Anwendung finden müssen, da auch bei großen Software-Entwicklungsprojekten die Phase der Konzeption und der Verläßlichkeitsprüfungen zu kurz bemessen wird und stattdessen nach einem informellen *bottom up*-Verfahren auf schnellstmögliche Implementation abgezielt wird (vgl. Spillner 1994).

Es ist naheliegend, nach der Rolle der Linguistik bei der Erstellung und Verarbeitung strukturierter Daten zu fragen, wenn man sich vergegenwärtigt, wie Alschuler SGML charakterisiert:

> If there is one single aspect that characterizes SGML [...] it is that it puts the computing power of information technology behind the all-encompassing descriptive power of human language. [Alschuler 1995, 1]

Der Kernpunkt dabei ist, wie bereits erwähnt, die Nutzung einer Grammatik für die Beschreibung der strukturellen Gesetzmäßigkeiten innerhalb der Daten. Natürlich wer-

den auch in der Informatik formale Grammatiken verwendet, in der Linguistik jedoch werden sie auf empirische Daten bezogen, so daß stets ein Wechselspiel zwischen formalen und empirischen Aspekten bei der Grammatikographie besteht.

Im einzelnen korrelieren die drei erwähnten Schritte in der folgenden Weise mit linguistischen Techniken:

1. Die strukturelle Analyse vorhandener Daten entspricht der Analyse eines Sprachkorpus, der als Grundlage für die grammatische Beschreibung einer Sprache genutzt werden soll. Für eine derartige Sprachdokumentation liegen Methodiken vor, die vor allem die Effizienz, den Abdeckungsgrad und den Umgang mit Korpuslücken betreffen.

2. Die empirische Exploration der zukünftigen Anforderungen an die Strukturierung entspricht in gewissen Aspekten der Feldforschung bzw. dem Informanteninterview. In diesem Bereich allerdings ist der Bedarf für weitere methodische Entwicklungen am größten. Dabei kann auf Erfahrungen im Bereich des *Knowledge Engineering* zurückgegriffen werden, wo man sich schon seit längerer Zeit der Probleme des geleiteten Erwerbs von Domänenwissen bewußt ist (s. z.B. Huber/Mandl 1982). Die Ermittlung des Wissens beispielsweise über die Struktur von Wörterbuchartikeln, das in einem Sachbuchverlag meistens fast ausschließlich in den Köpfen von Redakteuren und Autoren vorliegt, kann als ein Sonderfall des Wissenserwerbs interpretiert werden.

3. Die formale Spezifikation der Strukturbeschreibung muß mehreren Anforderungen gleichzeitig Genüge tun. Die DTD muß einerseits auf möglichst wenigen Kategorien, den sog. Elementen, beruhen und durchsichtig sein, zum anderen aber modular aufgebaut und parametrisierbar sein und die an sie gestellten Verarbeitungsansprüche erfüllen können. Solche Anforderungen stellen sich in der gleichen Weise bei der Spezifikation von natürlichsprachlichen Grammatiken für die maschinelle Sprachverarbeitung im Bereich der Computerlinguistik.

Damit zeigt sich das ungewöhnliche Bild, daß Methoden der Linguistik nicht nur indirekt, sondern unmittelbar übertragen werden können in einen anderen Anwendungsbereich und Linguistinnen und Linguisten somit prädestiniert erscheinen, diesen Transfer konkret zu leisten.

3 Digitale Texte – eine neue Themenstellung für die Linguistik

Im Lichte der vorangegangenen Ausführungen stellt sich die Linguistik als eine Wissenschaftsdisziplin dar, die sich mit der Struktur und der Verarbeitung eines hochkomplexen 'Datentyps' befaßt und dafür bestimmte Begriffe, Methoden und Techniken entwickelt hat, die auch für die Strukturierung und Verarbeitung anderer, meist weniger komplexer Datentypen herangezogen werden können. Vor allem Methoden der Grammatikographie lassen sich übertragen auf das Gebiet der strukturierten Repräsentation von Texten, aber auch nicht-sprachlicher oder nur teilweise sprachlicher Daten.

Auf ihrem bisherigen Entwicklungsweg scheint die Informationstechnologie bei der Sprache noch nicht wirklich angekommen zu sein. Das ist vielleicht auch auf den bislang vorherrschenden Ansatz der kognitiv adäquaten Simulation menschlicher Fähigkeiten zurückzuführen, der zwangsläufig große Rechenintensität und qualitative Unvollkommenheit mit sich bringt. Der Aufschwung des Client-Server-Prinzips auch auf der Ebene der Daten – exemplarisch belegbar mit dem *World Wide Web* – ermöglicht realistischere, erfolgversprechende Sprachverarbeitungskonzepte. Insgesamt ist es sicherlich nützlich, von dem Zweig der Sprachtechnologie, der inzwischen starken Eigencharakter besitzt und sich deutlich in eine ingenieurswissenschaftliche Richtung bewegt hat, so daß genuin linguistische Aspekte zunehmend in den Hintergrund treten, den Zweig der Texttechnologie mit seiner anwendungsbezogenen Umsetzung als *Textdesign* und *Hypertext Engineering* zu unterscheiden. Diese Bereiche, so wie sie im vorliegenden Buch skizziert werden, greifen dabei nicht nur auf das Gebiet der Textlinguistik und der Computerlinguistik zu, sondern machen auch von den Methoden anderer linguistischer Teilbereiche Gebrauch. Die strategischer Bedeutung für die Linguistik ist darin zu sehen, daß eine Ausweitung des anwendungsnahen Bereichs von der bislang vorherrschenden reinen Sprachtechnologie zur Texttechnologie bis hin zur Informationstechnologie im oben beschriebenen Sinne unmittelbare Anwendungsrelevanz aufweist und zugleich auf einem noch offenen Gebiet Maßstäbe definieren und Kompetenzfunktionen besetzen kann, wodurch die Chance gewahrt wird, die beginnende Entwicklung inhaltsbezogener Informationstechnologie mitgestalten zu können.

Literatur

Huber, G. L. und H. Mandl (1982): *Verbale Daten: Erhebung und Auswertung*. Weinheim: Beltz.

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

Sperberg-McQueen, Michael und Lou Burnard (Hrsg., 1994): *Guidelines for Electronic Text Encoding and Interchange*. Chicago, Oxford: Text Encoding Initiative [zwei Bände].

Spillner, A. (1994): „Kann eine Krise 25 Jahre dauern?". In *Informatik-Spektrum* 17, 48-52.

Travis, Brian E. und Dale C. Waldt (1995): *The SGML Implementation Guide. A Blueprint for SGML Migration*. Berlin: Springer.