Eine grammatikbasierte Integration von Hypertext und wissensbasierten Systemen

KLAUS PRÄTOR
Universität Düsseldorf

Für die folgenden Überlegungen gab es einen theoretischen und einen praktischen Anstoß. Der erste besteht in der These, daß Datenbanktheorie und Expertensysteme ein genuines Anwendungsfeld der Linguistik darstellen - und zwar nicht nur, wo es um natürlichsprachige Abfrage oder die Grammatiken der benutzten Programmier- oder Abfragesprachen geht. Der Grund ist ganz allgemein die Sprachförmigkeit der diesen Systemen immanenten Strukturen. (vgl. Prätor 1990) Obwohl diese These vor einem sprach philosophischen Hintergund nicht sonderlich überraschen kann, gibt es doch nur ein geringes Bewußtsein von ihr in den angesprochenen Anwendungsfeldern. Gleichwohl schlägt die Sprachförmigkeit auf die Begriffsbildung durch. Datensätze werden entweder als Objekte mit gewissen Attributen oder nach dem Muster der Prädikatenlogik als Sachverhalte in Bezug auf einen oder mehrere Gegenstände betrachtet.

Der praktische Anstoß resultiert aus der hohen Akzeptanz von Hypertextsystemen bei den Anwendern von wissensbasierten Systemen auf medizinischem Gebiet. Da deren Entwicklung mein gegenwärtiges Aufgabenfeld darstellt, kann mich das nicht gleichgültig lassen. In der Tat bilden Hypertextsysteme - wegen des geringeren Strukturierungsaufwands, der bequemen Handhabungsweise und der besseren Anschlußmöglichkeit an traditionelle Formen der Wissensdarstellung - häufig eine überlegenswerte Alternative zu wissensbasierten Systemen.

Die teilweise vielleicht etwas zu schematische Gegenüberstellung geht von der ursprünglichen Intention der Expertensysteme aus, einen Experten zu simulieren, was sich augenfällig im Stellen von vielen Fragen äußert, und kontrastiert sie mit dem Werkzeugcharakter der Hypertextsysteme, die die Führung ganz klar beim Benutzer belassen. Sie ordnet die Systeme, obwohl sie die jeweiligen Leistungen deutlich erweitern, den Grundtypen der Datenbanken mit hochstrukturierten Informationen und der Verarbeitung gering strukturierter Fließtexte zu. Daraus ergibt sich sowohl die Leistungsfahigkeit der Datenbanken für die Selektion und Umstrukturierung von Daten wie auch als Kehrseite ihr außerordentlich hoher Bedarf nach Standardisierung und Strukturierung der Informationen. Im Vergleich zu ihnen geben sich die Hypertexte mit einer flexibleren Strukturierung zufrieden. Sie ermöglichen so einen besseren Anschluß an die traditionellen Medien des wissenschaftlichen Informationsaustausches, besonders also die Druckwerke, und damit an die aktuelle fachliche Diskussion, häufig ein Schwachpunkt von Expertensystemen. In der abschließenden Charakterisierung der Wissensrepräsentation in Hypertextsystemen als menschenfreundlich ist die eigentliche Provokation, in diesem Zusammenhang überhaupt von Wissensrepräsentation zu sprechen. Nach dem Sprachgebrauch der Informatik liegt diese hier nicht vor. Man sollte aber nicht vergessen, daß in einem allgemeineren Verständnis Texte nicht nur einen Fall, sondern geradezu das Paradigma von Wissensrepräsentation darstellen, daß sie lediglich nicht den gegenwärtig realisierbaren Möglichkeiten automatischer Inferenz zugänglich sind. Man kann sogar umgekehrt behaupten, daß die Rede vom Übergang von Datenbanken zu Wissensbanken nur insofern Sinn macht, als darin die Annäherung der zunächst beliebigen Datenstrukturen an die semantischen Struktur von Sätzen, und damit von Textelementen, zum Ausdruck kommt.

Daß die Möglichkeiten der Hypertextsysteme im Hinblick auf die Behandlung von Wissen nicht so bescheiden sind, wie man auf den ersten Blick annehmen möchten, zeigt sich daran, daß man einfache Expertensysteme, nämlich solche mit determiniertem Suchbaum, durch Hypertextsysteme ersetzen kann. In der Abbildung 2 wird ein Ausschnitt aus einem kommerziell erwerbbaren System zur Leberdiagnostik in seiner Baumstruktur dargestellt, das in seiner Funktionalität vollständig durch ein Hypertextsystem ersetzt werden könnte.

Das Prinzip der Nachahmung ist einfach. An jedem Entscheidungsknoten ist das System auf Informationen angewiesen. Im Hypertextsystem werden dem Benutzer Fragen gestellt und er muß die der richtigen Antwort entsprechende Taste drücken und gelangt so entweder an die anschließende Frage oder, wenn der Entscheidungsbaum abgearbeitet ist, zur Antwort. Möglich ist das, wie bereits gesagt, nur bei einem relativ einfachen Typus von Expertensystemen. In komplizierteren Fällen ist die Ersetzung nicht mehr möglich.

Wünschenswert ist dann aber eine wechselseitige Ergänzung der Leistungen. Im Einzelfall können

Expertensystem	Charakter	**Hypertextsystem**
Subjektcharakter		Werkzeugcharakter
eher passiv	Benutzerrolle	eher aktiv
Regeln	charakteristische Datenstruktur	Link
Inferenz	charakteristische Operation	Bewegung in Text
Datenbank	Sytemtypus	Textverarbeitung
leistungsfähig	Schematisches Operieren	wenig ausgeprägt
sehr hoch	Strukturierungserfordernisse	geringer
schwierig	Anschluß an traditionelle Formen der Wissensdarstellung	gut
maschinenfreundlich	Wissensrepräsentation	menschenfreundlich

Abb. 1 Gegenüberstellung von Hypertext und Expertensystemen

dafür eine Vielzahl unterschiedlicher Gründe sprechen, von denen nur einige wichtige genannt werden sollen. Geht man von der Seite des Hypertextes aus, so kann Bedarf bestehen nach - einer punktuellen Unterstützung durch ein Expertensystem, z.B. zur Artbestimmung innerhalb eines biologischen Handbuchs. Vergleichbares ist für Reparaturanleitungen (Störfälle) oder medizinische Texte (Krankheitsbilder) vorstellbar. - zur Orientierung in komplexen Hypertextsystemen. Das lost in hyperspace" - Phänomen wird in fast jeder einschlägigen Veröffentlichung erwähnt. Auch für dieses Problem können expertensystemartige Unterstützungen eine Hilfe bieten.

Umgekehrt läßt sich ein Expertensystem ergänzen durch - die Möglichkeit der Erläuterung des Hintergrundes von Fragen des Systems oder allgemein von verwendeten Begriffen. So kann z.B. die Methodik eines durchzuführenden Tests erklärt werden. - die Verzweigung in Handbücher, Quellenangaben, Aufsätze als Bindeglied zu den geläufigen Formen der Fachkommunikation. - die Darstellung von im Expertensystem nicht oder schwer repräsentierbarem Wissen, z.B. der Ablauf von Stoffwechselvorgängen oder von anderen komplexen Abläufen. Durch grafische Darstellungen können topologische Zusammenhänge repräsentiert werden. - Möglichkeiten der Wissensakquisition, eine Ergänzung von ganz zentraler Bedeutung. Texte können in einer vorläufigen Strukturierung bereits der Benutzung zugänglich gemacht werden und durch weitere Erschließungs-schritte in die eigentlichen Expertensystemkompo-nenten übergeführt werden.

Diesem Wunsch nach Ergänzung steht die Tatsache entgegen, daß wissensbasierte und Hypertextsysteme ganz unterschiedliche Darstellungsweisen verkörpern. Beide Strukturtypen liegen sozusagen völlig windschief zueinander. Der Wunsch nach wechselseitiger Ergänzung beruht ja gerade darauf, daß man keinen einfachen Weg sieht, die Eigenheiten des einen Systems durch solche des anderen wiederzugeben. Man kann einwenden, daß einige Systeme existieren, die eine solche Ergänzung darstellen. So gibt es zum Beispiel eine Kopplung der Expertensystemshell Nexpert Object mit Hypercard und das System Knowledge Pro, das eine Integration beider Systemarten verspricht. Ohne die Leistungsfähigkeit dieser Systeme in Zweifel zu ziehen und ohne hier in Einzelheiten zu gehen, ist doch anzumerken, daß häufig entweder das Maß der Integration nur gering oder die Hypertextidee nur unzureichend verwirklicht ist, daß mithin die spezifischen Strukturen und Möglichkeiten textlicher Information zumindest teilweise ein-geschränkt werden. Insbesondere Hypercard und seine Nachahmer orientieren sich in ihrem Design mehr an der Datenbank- als an der Textmetaph0rik.

Relationen und wissensbasierte Systeme

Selbst wenn man zufriedenstellende Systeme zur Verfügung hat, kommt man, um diese vernünftig benutzen zu können, nicht umhin, sich selbst ein

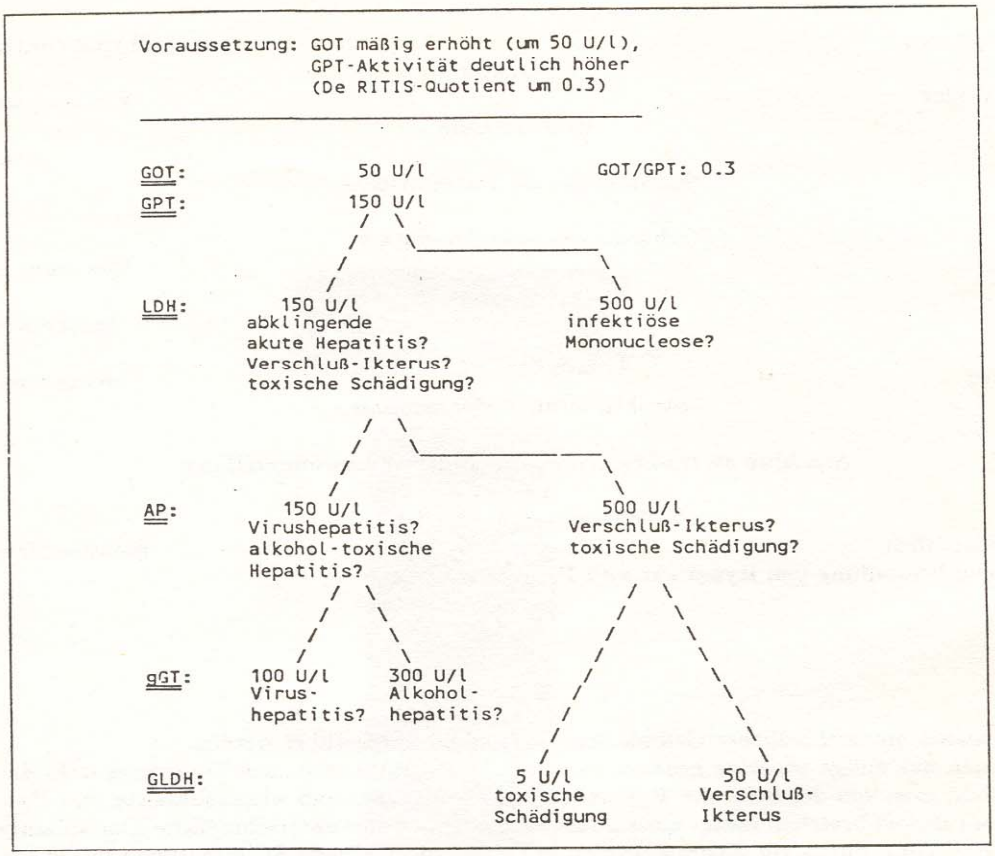

Abb. 2 Hypertext als Expertensystem

Bild von ihrem Zusammenspiel zu machen. Dazu muß erst ein gemeinsamer konzeptueller Nenner für ihre Integration gefunden werden. Als ein solcher wird hier ein relationaler Ansatz vorgeschlagen. Das Vorgehen besteht darin, zunächst zu fragen, wie man wissensbasierte Systeme und Hypertextsysteme in einem relationalen Modell darstellen kann und dann zu sehen, welche Möglichkeiten des Zusammenspiels beider Systemarten auf dieser Grundlage bestehen.

Ein relationaler Ansatz ist zur Zeit besonders im Bereich der Datenbanken vertraut, wo er derzeit den eindeutig dominierenden Typ darstellt. Sowohl die zu speichernden Objekte wie auch deren Beziehungen werden als Relationen abgelegt, die ersteren als Relationen ihrer Attribute. Die Darstellungsweise deckt sich völlig mit der in der Logikprogrammierung gebräuchlichen, deren Hauptvertreter gegenwärtig PROLOG darstellt. Die Prädikationen in der Prädikatenlogik stellen ein oder mehrstellige Relationen dar, die sich auch eins zu eins in relationale Datenbanken abbilden lassen.

Was die Logikprogrammierung und ganz generell auch die wissensbasierten Systeme von einfachen relationalen Datenbanken in technischer Hinsicht unterscheidet, ist die Verwendung von Regeln und der Einsatz von darauf beruhenden Schlußfolgerungen oder Inferenzen. Zweifellos stellen auch Regeln Relationen dar, nämlich Wenn-Dann-Beziehungen zwischen einem Antecedens und einem Succedens. Sie nehmen allerdings einen Sonderstatus ein, weil sie neben ihrer deklarativen Funktion auch einen prozeduralen oder operationalen Anteil haben. In ihrer wissensrepräsentierenden Funktion, d.h. in ihrem deklarativen Aspekt, lassen sich Regeln als Relationen darstellen. Das hat sogar den Vorteil, daß an die Stelle einer globalen Folgebeziehung inhaltlich qualifizierte Relationen treten können. Man kann dann beispielsweise unterscheiden zwischen Relationen, die Faustregeln bei der Lösung eines Problems darstellen, kausalen und semantischen Folgebeziehungen. Durch explizite Regeln wird dann nur noch bestimmt, wie der Schlußfolgerungsmechanismus das relational repräsentierte Wissen abarbeitet.

Relationen und Hypertext

Texte stellen eine völlig andere Datenstruktur dar. In der Regel werden sie in der Datenverarbeitung einfach als strings, als Zeichenketten wiedergegeben. Wählt man eine etwas höhere konzeptuelle

Pos	Wort		
1.1	Dies		
1.2	ist	100.1	Dies
1.3	der	100.2	ist
1.4	erste	100.3	eine
1.5	Satz	100.4	Erläuterung
1.6		100.5	zum
2.1	Dies	100.6	ersten
2.2	ist	100.7	Satz
2.3	der	100.8	
2.4	zweite		
2.5	Satz		
2.6			
3.1	Dies		
...			

Abb.3 Relationale Darstellung eines Textes

Ebene, so erscheinen sie als Listenstrukturen, als Listen von Wörtern etwa, oder auch in komplizierterer Form als Liste von Sätzen, die ihrerseits Wortlisten darstellen. Diese Struktur hat viele Vorteile. Sie eignet sich zum Beispiel gut zum Parsing. Für unserer Zwecke hat sie nur den Nachteil, von einer relationalen Darstellung, wie wir sie für die wissensbasierten Systeme erreicht haben, weit entfernt zu sein. Dabei ist es ein geringer Trost, daß Listen in KI-Sprachen wie LISP und PROLOG eigentlich als Relationen gehandhabt werden, nämlich als solche zwischen dem ersten Element der Liste und dem Rest. Auf dieser Basis wird die Gesamtliste rekursiv aufgebaut bzw. abgearbeitet. Dies ist aber eine sehr spezielle relationale Struktur mit der typischen Listeneigenschaft, nur element weise vom ersten Element her zugänglich zu sein. Demgegenüber hatten wir im Bereich der Datenbanken und wissensbasierten Systeme Gruppen von gleichförmigen, parallel angeordneten Relationen. Die Darstellung von Texten in dieser Form ist einfach und ausgefallen zugleich. In einer zweistelligen Relation werden die Wörter eines Textes und die jeweilige Position im Text aufgelistet, etwa wie das in Abbildung 3 dargestellt ist. Die Position könnte einfach durch fortlaufende Numerierung oder durch eine hierarchische Numerierung der Paragraphen, Sätze und Wörter erfolgen.

Diese Darstellungsweise hat hier nur den Status eines Denkmodells. Sie soll also nichts darüber besagen, wie eine hinreichend effektive Implementierung aussehen könnte. Versuche mit kleineren Texten in PROLOG, eine derartige Datei in gewohnter Gestalt am Bildschirm auszugeben, hatten allerdings recht gute Ergebnisse. Die Bezeichnung der Position kann schwierig werden, wenn Einfügungen und Streichungen sowie die Einbeziehung beliebig vieler und großer Dokumente möglich sein sollen. Hier existieren bereits Lösungen, wie sie z.B. von Nelson im Rahmen des Projektes XANADU ausgearbeitet wurden. (Nelson 1988) Mit diesen Methoden kann auch gleichermaßen auf einzelne Text

stellen (Wörter, evtl. auch Zeichen) zugegriffen werden, wie auch auf beliebige Textbereiche.

Ein angenehmer Nebeneffekt dieser Dateistruktur ergibt sich, wenn man den Index auf die zweite Spalte setzt, also statt nach den Positionen nach den Wörtern sortiert. Dann erhält man anstelle des fortlaufenden Textes den vollständigen Index des Dokuments, also die invertierte Datei.

Damit ist bisher nur die Umsetzung normaler Texte in die relationale Darstellung geklärt worden. Typische Hypertextbeziehungen sind entweder von assoziativer oder annotativer Art. Assoziative Beziehungen verknüpfen zwei Textstellen durch einen "link" in der Weise, daß von einer an die andere gesprungen werden kann. Annotative Verknüpfungen bringen beim "Anklicken" eines Textbereichs mit der Maus einen bisher unsichtbaren Textbereich dauerhaft oder zeitweise in den Blick. Beide Weisen können durch Relationen zwischen den Textadresssen, den Positionen der Wörter, repräsentiert werden, Eine assoziative Beziehung z.B. in prologähnlicher Schreibweise durch link (ausgangsadresse, zieladresse). Für eine annotative Beziehung ist die Basis gleichfalls eine zweistellige Relation zwischen Textstellen, nur muß in diesem Fall, nachdem der Sprung in den vorher unsichtbaren Textteil erfolgte, am Ende dieses Teils wieder in den Ursprungstext zurückgekehrt werden. Man kann die Verzweigung in den Hintergrundtext auch so auffassen, daß der Default- Wert für die nächste Textposition jeweils der fortlaufenden Numerierung entspricht, daß aber im Fall der Verzweigung statt des Defaultwertes ein anderer eingesetzt wird. (Wenn man hier bei Defaults an Frames denkt, ist man nicht auf der falschen Fährte)

Inhaltliche Strukturierung

Vergleichbar der relationalen Repräsentation von Regeln in wissensorientierten Systemen lassen sich auch die Relationen für die Hypertextlinks wie auch die verbundenen Textteile selbst inhaltlich qualifizieren. Ein Verweis auf das Literaturverzeichnis (Textrelation) kann als solcher gekennzeichnet werden und eine Fußnote oder eine Überschrift (Text teile) gleichfalls. Solche relativ formalen Textstrukturen können nahtlos in inhaltlichere übergehen, wenn etwa die Aspekte einer Literaturangabe (Autor, Titel,..) oder die festgelegten Teile einer bestimmten Vertragsform berücksichtigt werden. Selbstverständlich sind auch elaboriertere Qualifizierungen vorstellbar, aber es wird auch bei diesen Beispielen deutlich, daß durch die Aspektbildung Strukturen möglich werden, wie man sie von Datenbanken her gewohnt ist, daß der Übergang zu hochstrukturierten Datentypen mithin fließend ist.

Es wurde bereits angesprochen, daß sich bei dem vorgeschlagenen Modell der normalen Textdarstellung eine Anordnung in invertierter Form als Index und damit gleichzeitig als Wörterbuch des Dokuments gegenüberstellen läßt. Wie auf der Textseite eine inhaltliche Strukturierung durch zusätzliche Qualifizierungen denkbar ist, so können auf der Thesaurusseite Strukturierungen durch semantische Beziehungen der Wörter (Synonomie, Hyponomie,..) oder durch ihre Einordnung nach inhaltlichen oder funktionellen Gruppen erfolgen. Dies ist u. a. deshalb von Wichtigkeit, weil neben der Ergänzung durch Hypertexteigenschaften die Einbeziehung eines Thesaurus eines der wesentlichsten Desiderate wissensbasierter Systeme sein dürfte. Dies wird zwar nicht immer so gesehen, aber sobald mit unterschiedlichen Modulen und Anschlüssen an andersartige Informationsquellen gearbeitet wird, wird ein Thesaurus als Schnittstelle benötigt. Abgesehen davon ist in einem entwickelten Thesaurus ein großer Teil des über einen Welt ausschnitt vorhandenen Wissens inkorporiert.

Die für Dokument und Thesaurus unterschiedlichen Strukturen können auch übergreifend benutzt werden. So lassen sich beispielsweise Wörter durch ihre inhaltliche Einordnung im Thesaurus und zu gleich durch ihr Vorkommen in einer bestimmten inhaltlich oder funktional gekennzeichneten Textpassage charakterisieren, z.B. ein Symptom in einem Abschnitt zu einem bestimmten Schadstoff oder ein geografischer Name im Rahmen eines Literaturverweises. In jedem Fall gibt die Gesamtinformation hier mehr Information als die Summe ihrer Teile. Von Relevanz ist das in jedem Fall für die das Textretrieval, darüber hinaus aber auch der sprachlichen Analyse und der darüber möglichen Weiterverarbeitung von Wissensinhalten. Insgesamt eröffnet sich damit eine Perspektive, nicht nur wissensbasierte Systeme und Hypertext zu integrieren, sondern diese auch noch mit Thesaurus und Retrievalfunktionen verbinden zu können

Grammatiken für Textstrukturen

Die bisherigen Überlegungen zeigen, daß die Rede von der Unstrukturiertheit von Fließtexten nur aus dem Blickwinkel ihrer gegenwärtigen Behandlung in der automatischen Informationsverarbeitung richtig ist. Texte können über sehr ausgeprägte Strukturen verfügen, nur sind diese nicht gleichförmig genug, um sie mit den vergleichsweise groben Werk zeugen ihrer üblichen Computerhandhabung zu erfassen. Natürlich ist in vielen Bereichen ein Bewußtsein von diesen Strukturen vorhanden, beispielsweise im Bereich des Verlagswesens und Schriftsatzes. Dieser ist deshalb interessant, weil er eine Schnittstelle zwischen dem inhalt lichen und formalen Umgang mit Texten bildet. Hier haben sich Verfahren der Textauszeichnung entwickelt, Texte im Hinblick auf ihre grafische Gestaltung markieren. Dabei hat sich weitgehend die Praxis der inhaltlichen Auszeichnung durchgesetzt: Um Flexibilität bezüglich nachträglicher Änderung und mehrfacher Verwendung von Texten zu erreichen, werden nicht direkt grafische Attribute markiert, sondern funktionale oder inhaltliche. Dabei wird natürlich die konventionelle Textdarstellung verwendet und die Textcharakterisierungen werden als spezifische Zusatzzeichen in den Text eingebettet. Die Verfahren sind mittlerweile so elaboriert, daß man von Textauszeichnungssprachen sprechen kann, für deren Erzeugung und Prüfung auch formale Grammatiken existieren. Einen defacto Standard bildet heute die Standard Generalized Markup Language (SGML).

< *book* >
< *ti* > Organizational Burnout in Health Care Facilities
< *au* > Earl A. Simendinger < *deg* > Ph.D.
< *au* > Terence F. Moore
< *ehp* >< *no* > Chapter 1
< *cf* > Introduction to Organzational Burnout
..... Text ...< *fnr* > 1 < I > ... Text ...

< *h1* >Characteristics of Burned Out Organizations ... Text ...

< *h2* > Bickering
... Text... < *fn* >< *no* > 1 < *bb* >Health Service

Research
< *au* >Lloyd Connely< *au* > Dennis Pointer< *au* > Hirsh Ruschlin< *atl* >Viability and Hospital Failure: Methodology Considerations and Empirical Evidence< *7atl* >< *obi* >13 (Spring 1978): 27-36.< */fn*>
< */book* >

Abb. 5 Textauszeichnung mit SGML

Abbildung 5 zeigt markante Elemente des Dokumenttyps book. Die meisten Abkürzungen sind zu erraten:< ct > bezeichnet eine Kapitelüberschrift,< *fn* > eine Fußnote,< *fnr* > eine Referenz auf eine Fußnote, < *hl* > und< *h2* > eine Überschrift erster bzw. zweiter Ordnung, < *bb* > Bibliographie, < *atl* > Artikeltitel und < *obi*> weitere bibliographische Angaben. Das Ende einer Markierung wird durch< /... > gekennzeichnet oder ergibt sich eindeutig aus der folgenden Markierung.

Eben diese Textauszeichnungssprachen oder mark-up- languages sind gut geeignet, die im Zusammenhang eines relationalen Textmodells ange-

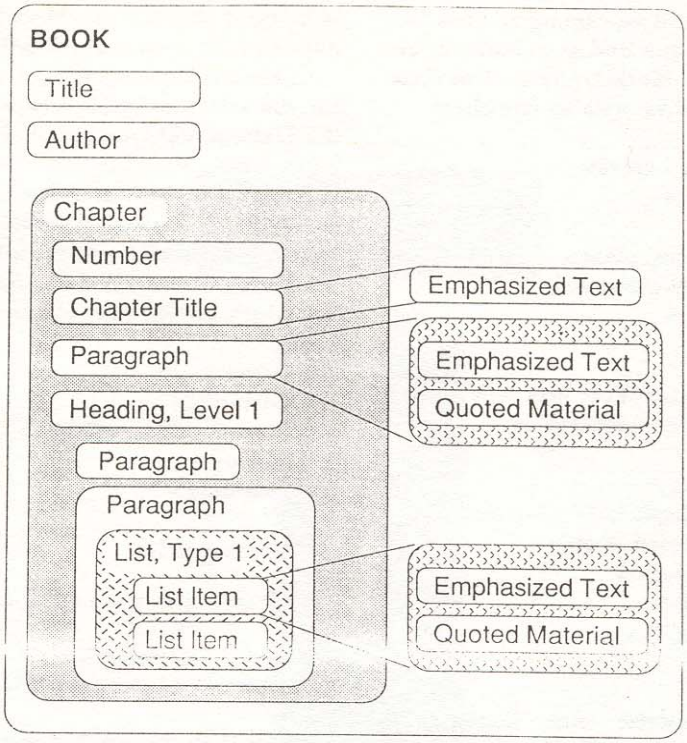

Abb. 4 Textstrukturen (nach Sperberg-McQueen1990)

sprochenen Qualifizierungen wiederzugeben. Zwischen beiden Modellen ist eine wechselseitige Abbildung möglich. Das heißt, daß auch diese Strukturierung flexibel hinsichtlich hoher und geringer Strukturierung ist. (Allerdings ist sie für hochstrukutrierte Daten nicht effizient.) Zu den üblichen Auszeichnungen gehören solche für hierarchisch geordnete Überschriften und die zugehörigen Textkörper sowie solche für Anmerkungen beziehungsweise Fußnoten. Beide können für die Umsetzung von annotativen Hypertextlinks benutzt werden Die jeweils untergeordnete Ebene kann relativ zur übergeordneten als zu verbergender Text betrachtet werden, der nach Belieben sichtbar gemacht und in sich natürlich wiederum verborgene Textteile enthalten kann. Anmerkungen würden in der Regel als temporär sichtbar werdende Textteile behandelt. Wo Auszeichnungsmöglichkeiten für Hypertexteigenschaften nicht standardmäßig vorhanden sind, lassen sie sich unschwierig hinzufügen. So schlägt die Text Encoding Initiative TEI (Sperberg-McQueen 1990) für assoziative Links, für Textsprünge also, eine Kodierung für Querverweise vor, die Einzeltextstellen und Textbereiche als Quelle und Ziel (auch in externen Dokumenten) handhabt, darüber hinaus Typisierung der Links (Qualifizierung) und die Verwaltung von Autor und Entstehungszeitpunkt der Links erlaubt. Textauszeichnungssprachen ebenso wie relationale Textstruktur erlauben die fast beliebige Ergänzung der Grundstrukturen durch jeweils für den Einzelfall zu definierende inhaltliche Kennzeichnungen.

Als ein weiterer Vorzug der mark-up languages erweist sich die Möglichkeit ihrer grammatikbasierten Erzeugung und Kontrolle. Abgesehen von der Nützlichkeit, die Korrektheit der Auszeichnung eines Textes zumindest formal kontrollieren zu können, eröffnen sich damit verbesserte Möglichkeiten der sprachlichen Analyse. An die Grammatik der Textstruktur kann sich eine in die in engeren Sinn sprachliche Struktur des Textes hineinreichende grammatische Analyse anschließen. Durch die Markierungen, besonders wenn sie inhaltlich angereichert sind, können zusätzliche Hilfen gegeben werden, die in vielen Fällen eine automatische Auswertung erst ermöglichen. Umgekehrt kann sprachliche Analyse auch die Auszeichnung der Texte erleichtern, indem sie, z.B. unterstützt von einem Thesaurus, nur in Zweifelsfällen eine explizite Markierung durch den Benutzer erforderlich macht.

Grundsätzlich hat eine Grammatik für Textauszeichnungssprachen zweierlei zu leisten. Zum einen hat sie die möglichen Ausprägungen eines Dokumenttyps zu kontrollieren und zum anderen die korrekte Anbringung der Markierungen. Das erste könnte im Fall eines Buches etwa so aussehen.

<book>→<title><author><text>
<text>→<chapter><text>
<text>→<chapter>
<chapter>→<chaptitle><chaptext>
<chaptext>→<paragraph><chaptext>
<chaptext>→<paragraph>

Eine inhaltlicher gefüllte Struktur ließe sich andeutungsweise folgendermaßen wiedergeben:

<Toxikologisches Handbuch>→
<Einleitung><Schadstoffkapitel> *[1]<Litverz>
<Schadstoffkapitel>→<Terminologie>
<allg.Stoffdaten><Toxikologie> .....
<Toxikologie>→<akute T.>
<chronische T.><spezielle T.>
...

Dabei wird üblicherweise eine kontextfreie Grammatik für die Erzeugung der Strukturen zugrundegelegt werden. Kontextfreie Sprachen erzeugen typischerweise lineare Zeichenketten. Eine interessante Alternative dazu bieten relationale Grammatiken (Heydthausen 1988, S. 73 ff.) Im Unterschied zu Chomsky- Grammatiken sind relationale Grammtiken nicht auf die Produktion linearer Zeichenketten, sondern auf die Generierung beliebiger Strukturen angelegt. Relationale Grammatiken erzeugen komplexe Gebilde durch die Beschreibung der Beziehungsstruktur ihrer Komponenten, die ihrerseits wieder komplex sein können. Interessant ist dieser Ansatz, weil er sich konzeptuell eng an den für die vorgetragenen Überlegungen zentralen Gedanken der relationalen Struktur anschließt, weil er semantischer orientiert ist als die Chomsky-Grammatiken und weil relationale Grammatiken auch die Erzeugung nichtlinearer Strukturen erlauben. Hypertexte sind aber per definitionem nichtlineare Datenstrukturen, auch wenn ihre einzelnen Sichtweisen jeweils linear darstellbar sind.

Grammatiken für hochstrukturierte Daten

Es klang bereits wiederholt an, daß die Möglichkeiten der Textmarkierung abgesehen von der Effizienz bis in Bereiche hochstrukturierter Informationen reichen, die üblicherweise mit relationalen Datenbanken, eventuell auch mit wissensbasierten Systemen gehandhabt werden. Da andererseits behauptet wurde, die Modelle der Textauszeichnungssprachen und der relationalen Textmodellierung seien ineinander abbildbar, liegt es nahe, auch für die relationalen Datenbanken eine Erzeugung der Datenstrukturen durch Grammatiken ins Auge zu fassen.

Dazu betrachtet man am besten eine hochstrukturierte Information in textlicher Darstellung, also etwa eine Tabelle mit physikalischen Daten oder ein Literaturverzeichnis. Diese würden mithilfe einer mark-up-Sprache etwas vergröbert so dargestellt:

<Litverzeichnis>
  <Liteintrag>
  <Autor><Titel><Biblio>
  </Liteintrag>
  <Liteintrag>
  .....
</Litverzeichnis>

Die Grammatik wäre etwa so geformt:

<Litverz>→<Liteintr> *
<Liteintr>→<Autor><Titel><Biblio>

Ersichtlich ist die Bedingung, daß so ein Übergang möglich ist, die Iteration von gleichartigen Elementen, also die Existenz einer listenförmigen Struktur. Es handelt sich also um einen Spezialfall einer textlichen Struktur. Ob sich eine listenförmige Struktur ergibt, hängt aber auch von der Größe und Art der gewählten Einheiten ab, von der Auflösung sozusagen. Wählt man als Grenzfall Wörter, Sätze oder Abschnitte so lassen sich leicht Listen und damit relational darstellbare Strukturen finden. Dies war der Trick bei der hier vorgeschlagenen relationalen Textdarstellung. Mit der Grammatik für das Literaturverzeichnis läßt sich auch eine Tabelle einer relationalen Datenbank darstellen, wobei die zweite Regel jeweils genau einen Datensatz erzeugen würde. In diesem Fall wird die interne Struktur einer Relation mithilfe der Grammatik abgeleitet. Es ist aber auch möglich, in ähnlicher Weise das Relationsgefüge einer Datenbank zu erzeugen. Ausgehend von der grammatischen Skizze einer Buchstruktur

<Buch>→<Titel><Kapitel> * <Litliste>
<Kapitel>→<Abschnitt> *
<Litliste>→<Liteintrag>
<Liteintrag>→<Autor><Titel><Biblio>

würde ein Gefüge von drei Tabellen (oder Listen von Datensätzen) entstehen, die aufeinander in der Weise bezogen sind, wie dies durch die Grammatik vorgegeben ist. (Abb. 6)

Hierarchien von textlichen Listen werden durch Hierarchien von Tabellen wiedergegeben. Dabei ist

[1] * ist eine Abkürzung für Wiederholbarkeit

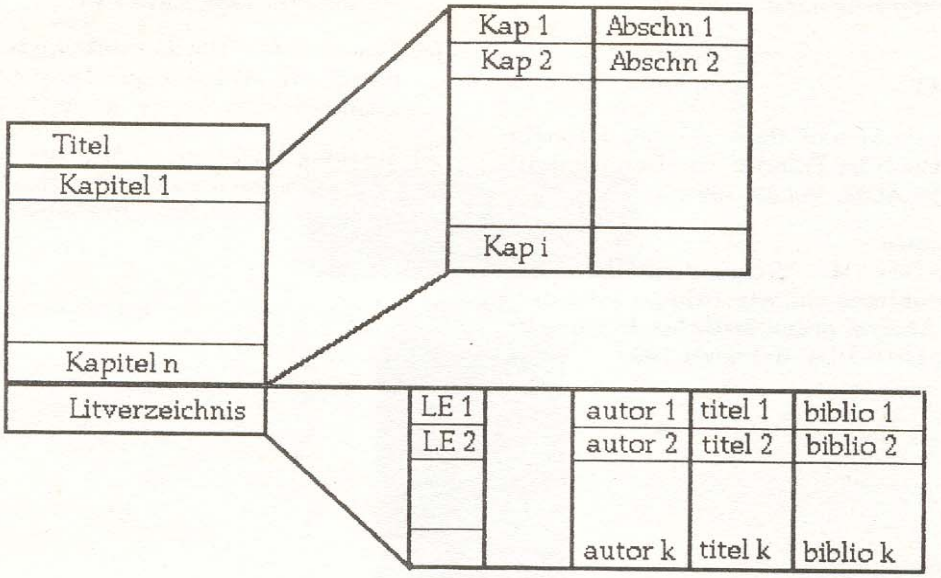

Abb. 6 Durch Grammatik erzeugte Dateistruktur

die Erzeugung der Tabellen mit Literatureinträgen (LE) nichtterminal. Die Datei wird nicht wirklich erzeugt. Sie ist nur ein Schritt auf dem Weg der grammatischen Produktion. Man sieht auch leicht, wie sich typische Hypertextstrukturen realisieren lassen: In sich verschachtelte Annotationen - Gliederungsstrukturen - bilden sich ab in eine Hierarchien von Relationen, die jeweils zwei Ebenen miteinander verknüpfen. Die so erzeugten Strukturen entsprechen teilweise denen, die in den Theorien der Normalisierung und Schlüsselbildung von Datenbanken oder auch in logisch-semantischen Datenbankmodellen behandelt werden. Neben dem konzeptionellen Brückenschlag ist dies ein zweites Argument, den in diesem Bereich wenig üblichen Gedanken der grammatischen Erzeugung weiterzuverfolgen. Ein weiterer Grund kommt hinzu: Datenstrukturen in dieser Perspektive zu betrachten, kann dazu beitragen, semantisch-syntaktische Aspekte stärker zu beachten und so den Übergang von einem datenorientierten zu einem wissensorientierten Ansatz zu befördern. Wissensbasierte Hypertextsysteme Die Übertragung des Gedankens der grammatischen Erzeugung auf hochstrukturierte Datensysteme findet ihre beste Entsprechung in der Frameidee, die eng mit semantischen Netzen und case grammars verwandt ist. Dieses Konzept unterscheidet sich nicht so sehr technisch von üblichen Datenobjekten - obwohl Differenzen in Form von Facetten und Triggern vorhanden sind -, sondern durch die Verknüpfung mit Konzepten der Wissensverarbeitung. Meines Erachtens ist es die semantisch-pragmatische Orientierung bei der Betrachtung der Funktionalität der einzelnen slots im Hinblick auf den Gesamtframe, die den wesentlichen Schritt dieses Konzepts hin zur Wissensverarbeitung ausmacht.

Betrachtet man ein Hypertextsystem in der referierten relationenartigen Darstellung aus dem Blickwinkel der Frameidee so erscheinen die textlichen Teile als besondere Slots. Diese können durch andere ergänzt werden, die den mithilfe der Textauszeichnungssprache angebrachten Zusatzqualifikationen entsprechen. Auf diese Typisierungen und auf die im Thesaurus niedergelegten Relationen stützen sich die Inferenzmöglichkeiten. Damit ist eine einheitliche konzeptuelle Darstellung der Wissenrepräsentation in Hypertext und wissensbasierten Systemen möglich. Mit geeigneten Werkzeugen läßt sich so ein integriertes System mit wissensbasierten und Hypertext-Elementen entwickeln. Die eigentlichen textlichen Teile sind nur für den menschlichen Benutzer interessant. Wegen ihrer geringen Strukturiertheit sind sie für die automatische Schlußfolgerung nicht zugänglich. Da sie aber in der vorgeschlagenen Strukturierung syntaktischer und lexikalischer Analyse gut zugänglich sind, besteht die Möglichkeit, ein derartiges System in Richtung auf höhere Strukturierung weiterzuentwickeln. Die zunehmende Aufbrechung textlicher Strukturen kann dabei als eine Form der Wissensakquisition betrachtet werden, die dem

Anschluß an die etablierten Formen der Wissensdarstellung entgegenkommt.

Literatur

[1] Carlson, D.A. und Ram, S., "HyperIntelligence: The Next Frontier", in: Communications of the ACM, Vol.33, 1990/3, S.311-321

[2] Heydthausen 1988
Heydthausen, M., "Diagnostische Sprache - Datenstrukturen und Algorithmen zur semantischen Analyse primärärztlicher Diagnosebezeichnungen," Diss. Hannover 1988

[3] Nelson, T. "Managing Immense Storage", in: BYTE Jan. 1988 S.225 - 242

[4] Prätor, K.M. "Die Sprachförmigkeit des Computers", in: Mitteilungen des Deutschen Germanistenverbandes, Jg. 37, 1990/3, S.15-19

[5] Sperberg-McQueen, C.M. und Burnard, L. (Hrsg.) "Guidelines For the Encoding and Interchange of Machine- Readable Texts", Chicago