Inhaltsorientierte Navigation in automatisch generierten Hypertext-Basen[1]

Udo Hahn Rainer Hammwöhner Ulrich Reimer Ulrich Thiel

Universität Freiburg Universität Konstanz GMD-IPSI

Linguistische Informatik Informationswissenschaft Darmstadt

Computerlinguistik

Zusammenfassung

Der automatische Aufbau von Hypertexten aus Kollektionen linearer Texte erfordert Verfahren zur Analyse und Segmentierung von Texten, sowie zur Generierung von Hypertext-Kanten. In diesem Beitrag werden Theorien zur Beschreibung der thematischen Struktur von Texten aufgegriffen und zur Entwicklung von Kriterien genutzt, die es erlauben, inhaltlich begründete Kanten zwischen Textfragmenten zu erzeugen. *Textgraphen* resultieren als netzwerkartige Repräsentationen von Texten und dienen einem auf objektorientierter Interaktion basierenden Dialogmodell als Ausgangsbasis zur semantisch kontrollierten Exploration der *Hypertext-Basis.*

1. Einleitung

Entsprechend den bislang vorherrschenden Anwendungsklassen für Hypertextsysteme - Ideenexploration und -Verwaltung *(idea processing)* und Autorenunterstützung (für technische Dokumente, Programmtexte u.a.) - setzen die Autoren beim inkrementellen Prozeß der Generierung und Modifikation eines Hypertexts die Kanten *manuell.* Will man aber die Funktionalität von Hypertextsystemen auf schon existierende Textkollektionen - etwa für die Zwecke des Information Retrieval, der Faktenextraktion, der Wissensexploration durch Browsing o.a.[2] - übertragen, stellt sich das Problem, Verfahren bereitzustellen, mit denen umfangreiche Volltexte bzw. Kollektionen von Volltexten in Hypertexte überführt werden können. Dabei ist zunächst eine Dekomposition der Originaltexte in eigenständige Texteinheiten, den Knoten des aufzubauenden Hypertexts, zu leisten, die dann durch inhaltlich begründete Kanten verbunden werden müssen. Für kleinere bis mittlere Textkörper ist dieser Vorgang noch intellektuell zu kontrollieren. Frisse (Frisse 88) schlägt z.B. ein semi-automatisches Verfahren vor, das die Zerlegung des Textes anhand oberflächensyntaktischer Indikatoren, z.B. Kapitelgrenzen, vorsieht, die inhaltliche Strukturierung des Hypertexts aber den Autoren überläßt. Große, über eine Vielzahl von Dokumenten sich erstreckende Hypertexte, wie sie insbesondere bei Anwendungen im Bereich des Information Retrieval zu erwarten sind, sind aber in ihren thematischen Interdependenzen nicht mehr intellektuell zu erfassen, so daß eine Automatisierung auch der Generierung von Hypertext-Kanten erforderlich wird. Kriterien für die Generierung von Kanten können aus der thematischen Struktur des Ausgangmaterials abgeleitet werden.

---

[1] Dieser Text ist erschienen in: Gloor, P.A. / Streitz, N.A. (Hrsg.) Hypertext und Hypermedia. Von theoretischen Konzepten zur praktischen Anwendung. Springer, 1990, S. 205-219.

2 Exemplarisch ist diese Konzeption für hochentwickelte Text-Interaktionssysteme bereits in Weyer/Borning 85 und Lenat et al. 83 am Beispiel elektronischer Enzyklopädien beschrieben worden.

Der entstehende Hypertext erweist sich als ein geeignetes Medium für explorative Retrievaldialoge (Bates 86), wie sie in Reaktion auf die Kritik am *Matching-Paradigma*[3] (Robertson 80) des konventionellen Information-Retrieval vorgeschlagen wurden. Darüber hinaus bilden Cluster relationierter Textsegmente eine Grundlage für kontextorientierte Relevanzmaße, die die Relevanz eines Textes nicht nur in Bezug auf ein Interessenprofil, sondern auch anhand der bereits zuvor präsentierten Textabschnitte bestimmen (Tiamyu/Ajiferuke 88).

Ein aus ursprünglich nicht aufeinander bezogenen Einzeltexten aufgebauter Hypertext erfordert eine besondere Unterstützung des Benutzers bei der Hypertext-Navigation. Im Gegensatz zu konventionellen, intellektuell erstellten Hypertexten sind die Kanten nicht absichtsvoll von Autoren bzw. Autorenkollektiven eingefügt, sondern rein thematisch motiviert. Zudem sind Inhalt und Umfang der Textkollektion auch von zufälligen Einflußgrößen abhängig, wie der Verfügbarkeit von Dokumenten, den zeitweiligen Interessen einer Benutzergruppe usf. Das führt dazu, daß wichtige Informationen fehlen, andere hingegen stark redundant repräsentiert sein können. Metainformation, die einzelne Texteinheiten in einen größeren Kontext einordnet, ist nicht vorgegeben, sondern muß während der Navigation erzeugt werden. Diese Problematik motiviert einen Navigationsstil für Hypertexte im Information Retrieval, der konversationale Aspekte der Interaktion mit objektorientierter graphischer Manipulation verbindet (Thiel 89, Thiel/Hammwöhner 89).

In diesem Kontext ist die Forschung im Rahmen der Projekte TOPIC[4] und TWRM[5]-TOPOGRAPHIC[6] angesiedelt. Während das Textkondensierungssystem TOPIC (Hahn/Reimer 86, 88) die inhaltliche Analyse von Volltexten[7] und den Aufbau von Repräsentationen der Textinhalte, sogenannten Textgraphen, leistet, wird der graphisch-interaktive Zugriff auf die aus Volltexten und ihren Repräsentationen aufgebauten *Hypertext-Basen* durch das wissensbasierte Retrievalsystem TWRM-TOPOGRAPHIC (Hammwöhner/Thiel 87, Kuhlen et al. 89) realisiert. Da eine vollständige Beschreibung des Gesamtsystems den Rahmen dieses Beitrags bei weitem sprengen würde, werden wir uns auf folgende Aspekte konzentrieren:

- Die Struktur der automatisch generierten Hypertexte wird in Kapitel 2 beschrieben, während auf die eigentliche Textanalyse und Textgraphgenerierung nicht eingegangen wird.
- Kapitel 3 führt ein Interaktionsmodell ein, das einen flexiblen Umgang mit derartigen Hypertexten erlaubt.

2. Der Textgraph als Hypertext

Das Ergebnis der Analyse und anschließenden Kondensierung eines Textes durch das TOPIC-System (Hahn/Reimer 88) ist ein Hypertext, den wir *Textgraph* nennen. Die Textgraphen mehrerer Texte bilden eine *Hypertext-Basis,* wobei jedoch mehrere Textgraphen durchaus

---

[3] Das *Matching-Pardigma* bezieht sich auf den kontextfreien Abgleich einer vollständig vorformulierten Suchanfrage mit einer Textmenge.

[4] TOPIC: **T**ext **O**riented **P**rocedures for **I**nformation Management and **C**ondensation of Expository Texts.

[5] TWRM: **T**ext**w**issens-**R**ezeptions-**M**echanismus

[6] TOPOGRAPHIC: **Top**ic **O**perating with **Graph**ical **I**nteraction **C**omponents

[7] Das momentan repräsentierte Sprach- und Weltwissen ist auf Produktbeschreibungen von Mikrocomputersystemen ausgerichtet. Um in den Beispielen die ständig wiederholte Nennung von Produkten der Firmen mit drei Buchstaben zu vermeiden, wurde der Phantasierechner *Zenon-X* kreiert.

einen *textübergreifenden* Hypertext bilden können. Die Knoten in einem Textgraphen sind durch verschiedene Kantentypen miteinander verknüpft, die alle automatisch generiert werden (Reimer/Hahn 88). Fünf Klassen von Hypertext-Kanten werden derzeit unterstützt:

1) **Abstraktionsrelationen:** Hierunter fallen Beziehungen, die eine Konzeptspezialisierung anzeigen, sowie framespezifische Aggregationsbeziehungen vom Typ 'Slot' und 'Eintrag', die unterschiedlich detaillierte *Zusammenfassungen* eines Textes miteinander verbinden.

2) **Relationen zwischen Themenbeschreibungen und assoziierten Textpassagen:** Kanten diesen Typs erlauben den Zugriff von einer Themenbeschreibung auf die Passagen des Originaltextes zu diesem Thema *(Text-Retrieval).*

3) **Relationen zwischen Themenbeschreibungen und assoziierten Fakten in der Textwissensbasis:** Einem Text entnommene Aussagen zu einem Konzept können direkt aus einer Themenbeschreibung heraus, in der dieses Konzept verwendet wird, zugegriffen werden *(Fakten-Retrieval).*

4) **Rekonstruierte Kohärenzrelationen:** Thematische Progressionsmuster der Themenentwicklung von Volltexten dienen als inhaltlich motivierte Tourenvorschläge für das Navigieren in Hypertexten.

5) **Konstruierte Kohärenzrelationen:** Aus der in Textgraphen repräsentierten referentiellen und semantischen Struktur von Texten werden automatisch intra- und intertextuelle Relationen abgeleitet, die dem Zweck dienen, aus unterschiedlichen Dokumenten entnommene Passagen, die in Bezug auf ein Diskursthema ergänzende, vertiefende oder kontrastierende Informationen enthalten, zu einem Hypertextpfad zusammenzuführen.

In den folgenden Abschnitten gehen wir besonders auf die unter den Punkten 1), 4) und 5) erwähnten Kantentypen näher ein.

2.1 Themenbeschreibungen und Abstraktionsrelationen

Ein Textgraph (Abbildung l zeigt einen Ausschnitt eines stark vereinfachten Textgraphen) leistet in erster Linie die *Repräsentation der thematischen Struktur* des zugehörigen Textes auf unterschiedlichen Konkretionsebenen gleichzeitig. Die Blattknoten eines Textgraphen stellen dabei die detaillierteste Untergliederung des Textes in thematisch kohärente Textpassagen dar, während die nicht-terminalen Knoten mehrere solcher Passagen mit ähnlichen Themen zu einer Passage generelleren Themas zusammenfassen. Steigt man einen Textgraphen von den Blattknoten her auf, ergibt sich somit eine Untergliederung des zugehörigen Textes in immer weniger Passagen zunehmend allgemeinerer Themen. Die Wurzelknoten geben schließlich die allgemeinste Charakterisierung des zugehörigen Textes als Ganzes an.

Auf den nicht-terminalen Ebenen eines Textgraphen wird die ursprüngliche, lineare Anordnung der einzelnen Passagen im Text aufgehoben, da ihre Zusammenfassung zu längeren Passagen nach inhaltlichen Kriterien erfolgt und nicht nach ihrer Anordnung im Text. Beispielsweise faßt der Knoten 2 in Abbildung l die Textpassagen 1-3 sowie 5-6 zusammen und läßt die sich dazwischen befindliche Textpassage 4 aus.

Neben der Zusammenführung verschiedener, inhaltlich ähnlicher Textpassagen übernehmen die Textgraphknoten auch die Funktion der inhaltlichen Beschreibung der ihnen zugehörigen Passagen. Dazu enthält jeder Knoten als *Themenbeschreibung* ein semantisches Netz. Im einfachsten Fall besteht es nur l aus einem Knoten, der den Namen des Konzepts angibt, von dem die betreffende Textpassage primär handelt. So ist das Thema der dem

Knoten 4 in Abbildung l zugeordneten Passagen (das sind in diesem Fall alle sechs) der Zenon-X, während die Passagen 1-3 und 5-6 von Herstellern (Knoten 2) handeln. Detailliertere Themenbeschreibungen geben zusätzlich zu den Konzepten, mit denen sich ein Textabschnitt befaßt, auch den Aspekt an, unter welchem das Konzept näher behandelt wird. In einem solchen Fall ist in der Themenbeschreibung eine Slot-Kante[8] vorgesehen, über die entweder die näher behandelte Eigenschaft des Konzepts angegeben ist, oder ein anderes Konzept, das mit dem ersten Konzept in einer inhaltlichen Beziehung steht und in bezug darauf ausführlicher diskutiert wird. Beispielsweise faßt der Knoten 5 in Abbildung l alle Textpassagen zusammen, die die Cpu des Zenon-X behandeln. Die Angabe eines Konzepts oder einer Eigenschaft mittels einer Slot-Kante ist kontextspezifisch. Das bedeutet, daß es in der betreffenden Textpassage nicht um dieses Konzept (z.B. Cpu im Knoten 5) im allgemeinen geht, sondern daß es nur in dem durch das übergeordnete Konzept gegebenen Kontext (z.B. Zenon-X im Knoten 5) behandelt wird.

Ist der durch eine Slot-Kante spezifizierte Aspekt selber nochmals detaillierter im Text behandelt, so wird mittels einer Eintragskante eine zusätzliche Charakterisierung in der Themenbeschreibung vorgesehen. Beispielsweise repräsentiert der Knoten 6 des Textgraphen in Abbildung l, daß die Textpassagen 1-3 nicht nur von den Produkten der Zeta-Maschinen GmbH handeln, sondern sich speziell mit einem bestimmten Datenbanksystem befassen.

Neben den Slot- und Eintragskanten können in einer Themenbeschreibung auch Kanten auftreten, die eine Konzeptspezialisierung anzeigen (vom Typ 'Is-a' und 'Instanz': zur Semantik siehe Reimer 89). Ihnen kommt im Textgraph jedoch nur eine Informierungsfunktion zu, indem sie verschiedene Konzepte mit gemeinsamen Oberbegriff durch Relationierung mit diesem Oberbegriff zu Gruppen zusammenfassen. Eine thematische Charakterisierung erfolgt durch solche Kanten nicht. Gleichwohl bieten sich damit beim Retrieval verschieden detaillierte Einstiegspunkte an.

Zusammenfassend wollen wir festhalten, daß die Themenbeschreibungsknoten in einem Textgraph keine Textknoten sind, sondern eine inhaltliche Beschreibung der ihnen zugeordneten Textpassagen bereitstellen (vgl. mit den 'Toc Nodes' in Trigg/Weiser 86). Nur diese Textpassagen bilden Textknoten; sie sind an den Blättern eines Textgraphen eingebunden (in Abb. l die Knoten mit der Bezeichnung 'Textpassage i).

Zwischen den Themenbeschreibungsknoten eines Textgraphen können Beziehungen unterschiedlicher Art bestehen, die durch entsprechende Hypertextkanten dargestellt sind:

- Eine *Identitätskante* besteht zwischen einem übergeordneten Knoten *n* und einem untergeordneten Knoten *n′* genau dann, wenn das Themenbeschreibungsnetz im Knoten *n* (eventuell nur als Teilgraph) auch im Knoten *n′* auftritt, jedoch nicht unterhalb einer Slot- oder Eintragskante, da die Angaben unterhalb solcher Kanten nur kontextspezifische Gültigkeit haben.
- Eine *Is-a-* oder *Instanzkante* besteht zwischen einem übergeordneten Knoten *n* und einem untergeordneten Knoten *n′* genau dann, wenn *n* nur einen Knoten enthält, und dieser ein Konzept bezeichnet, das ein (Is-a- oder Instanz-) Oberbegriff des obersten Knotens eines der Netze in *n′* ist.
- Eine *Slot-Kante* besteht zwischen einem übergeordneten Knoten *n* und einem untergeordneten Knoten *n′* genau dann, wenn *n* nur einen Knoten enthält, der auch in *n′* auftritt und dort durch eine Slot-Kante näher charakterisiert wird.

[8] Der Name des Kantentyps 'Slot' (und auch des weiter unten eingeführten Kantentyps 'Eintrag') leitet sich daraus ab, daß der Textanalyse und -kondensierung, die zum Aufbau eines Textgraphen führen, ein Frame-Repräsentationsmodell (Reimer 89) zugrunde liegt. Daraus ergibt sich auch die Semantik dieser Kantentypen.

- Eine *Eintragskante* besteht zwischen einem übergeordneten Knoten *n* und einem untergeordneten Knoten *n'* genau dann, wenn *n* ein Netz enthält, das aus zwei durch eine Slot-Kante verbundenen Knoten besteht, und in *n'* das gleiche Netz auftritt, dort jedoch um eine Eintragskante erweitert.

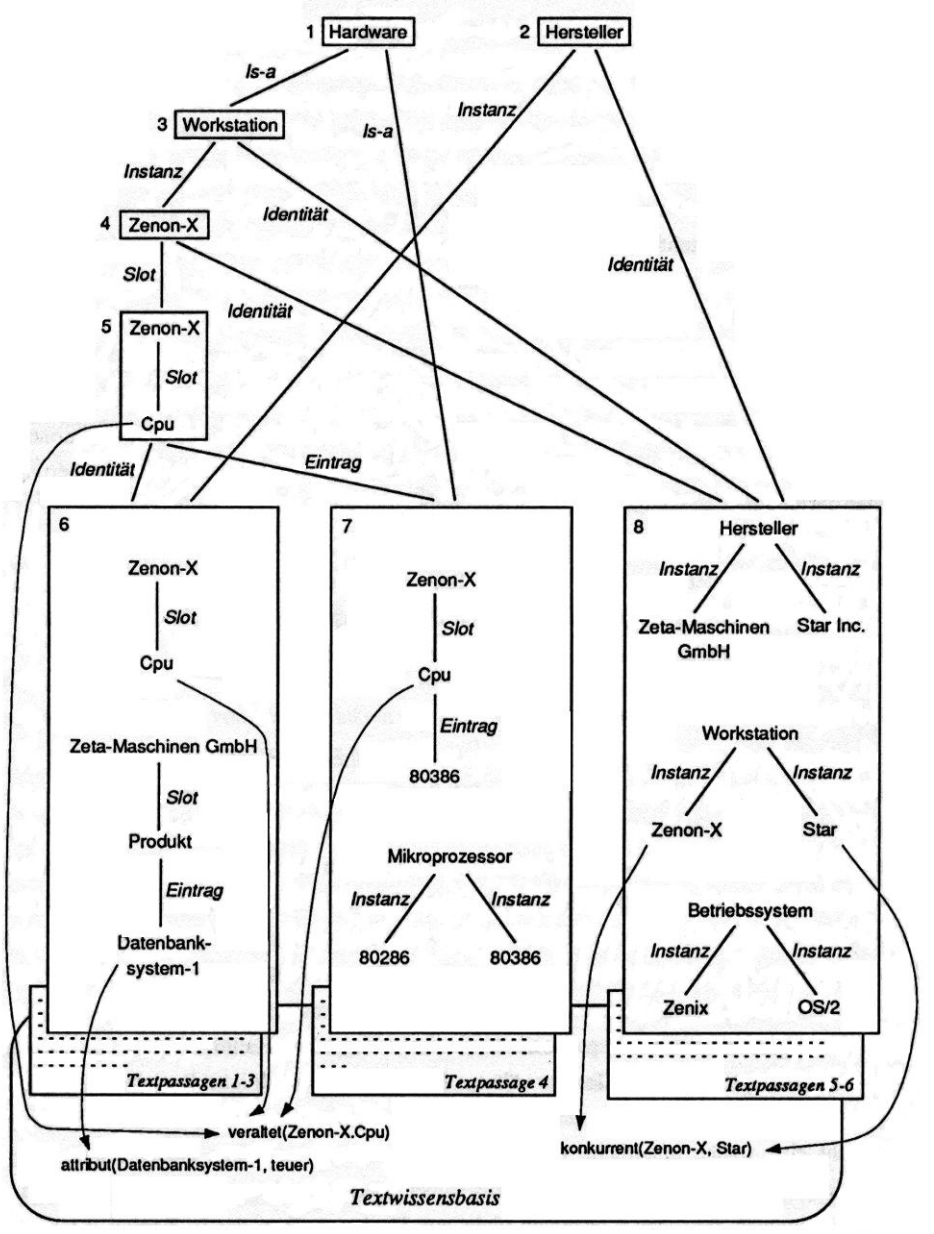


Abbildung 1 Ausschnitt eines vereinfachten Textgraphen (ohne Kohärenzrelationen)

Jeder der oben aufgeführten Kantentypen steht für eine Abstraktionsrelation zwischen zwei Themenbeschreibungen. Technisch betrachtet stellt die hierarchische Textgraphstruktur neben unmittelbaren *link point / link region*-Referenzen Ketten von *link point/link* point-Referenzen (wie sie in traditionellen Hypertextsystemen nicht auftreten; vgl. etwa Conklin 87) wachsender konzeptueller Spezialisierung bereit. Diese Ketten terminieren schließlich in einem *link point* (ein Blatt des Textgraphen).

Zusätzlich zur Verknüpfung durch Abstraktionsrelationen sind alle Textgraphknoten mit den assoziierten Textpassagen verkettet. Ferner ist aus jedem Textgraphknoten heraus, dessen Themenbeschreibung ein Konzept *k* verwendet, der Zugriff auf Aussagen über dieses Konzept möglich, die aus dem betreffenden Text gewonnenen wurden und in der Textwissensbasis abgelegt sind (vgl. Abb.l).

2.2 Rekonstruierte Kohärenzrelationen

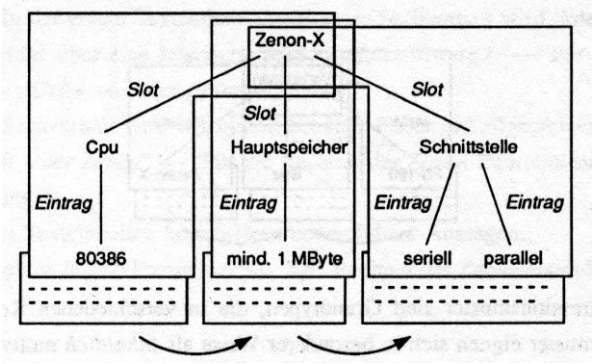
Orthogonal zu den im vorangegangenen Abschnitt behandelten Themenbeschreibungen ist die thematische Strukturierung eines Textes durch Kohärenzrelationen. Während die Themenbeschreibungen einen Text danach untergliedern, welche Konzepte in bestimmten Textabschnitten eine dominante Rolle spielen und wie diese Konzepte in einen allgemeineren thematischen Kontext eingebettet sind (durch Slot-und Eintragskanten dargestellt), liegen den Kohärenzrelationen bestimmte Verlaufsmuster der Konzepterwähnung zugrunde. Eine bestimmte Klasse solcher Verlaufsmuster, im folgenden *(thematische) Progressionsmuster* genannt, lassen sich auf lokale, textuelle Konnektivität stiftende Kohäsionsphänomene zurückführen. Dies sind in erster Linie Koreferenzbeziehungen zwischen sprachlich verschieden realisierten Erwähnungen derselben Konzepte sowie lexikalische Kohäsion (Halliday/Hasan 76). Lexikalische Kohäsion ist gegeben, wenn ein im Text erwähntes Konzept eine inhaltliche Beziehung zu einem vorher angesprochenen Konzept aufweist. Legt man eine Frame-Repräsentation von Konzepten zugrunde, dann entspricht eine solche inhaltliche Beziehung der Beziehung zwischen einem Frame und seinen Slots bzw. seinen Slot-Einträgen (die dann jeweils selber wieder für Konzepte stehen). Insofern spielen inhaltliche Beziehungen zwischen Konzepten nicht nur für den Aufbau der in Kapitel 2.1 diskutierten Themenbeschreibungen eine Rolle, sondern auch für die Bestimmung von Progressionsmustern, nur werden sie hierfür anders ausgewertet. Die Progressionsmuster treten sowohl innerhalb eines Absatzes als auch absatzübergreifend auf.

Drei Haupttypen von Progressionsmustern lassen sich unterscheiden (Daneš 74):

- Ein *konstantes Thema* liegt vor, wenn ein Konzept (das Thema) im Text eingeführt und anschließend in mehreren seiner Aspekte näher ausgeführt wird. Beispielsweise liegt dem folgenden Textausschnitt das konstante Thema 'Zenon-X' zugrunde:

  "Der Zenon-X ist mit einem 80386-Prozessor als Cpu ausgestattet. Dieser Prozessor ... Der Rechner wird mit mindestens l MByte Hauptspeicher geliefert. Er ist ausbaufähig bis ... Neben einer seriellen Schnittstelle verfugt der Zenon-X auch über einen parallelen Anschluß ..."

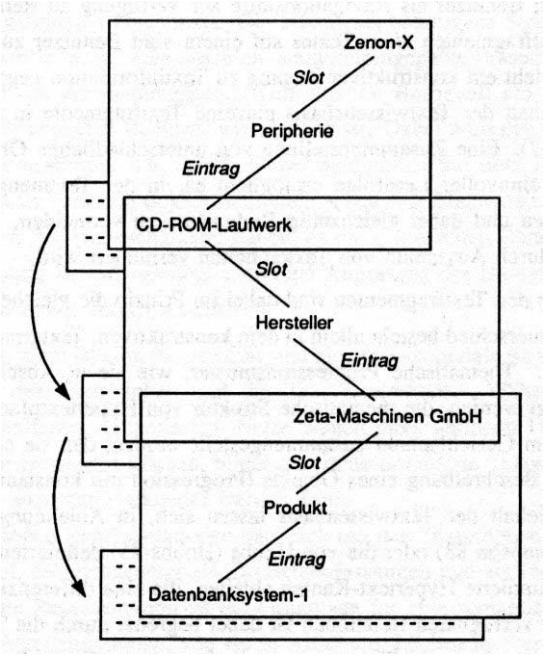
  Dieses Progressionsmuster läßt sich durch den untenstehenden Kohärenzgraphen darstellen. Dabei enthalten die einzelnen Knoten wiederum Themenbeschreibungen (wie in Abb.l), und ihnen ist jeweils als Textknoten die zugehörige Textpassage zugeordnet. Die Verknüpfung der Knoten spiegelt jetzt jedoch die Abfolge der Themen im Text wider:



- Eine *lineare Thematisierung von Rhemata* liegt vor, wenn ein zu einem Konzept (dem Thema) ausgeführter Aspekt (das Rhema) anschließend zum Thema wird. Im Beispiel:

  "Für den Zenon-X ist ein CD-ROM-Laufwerk mit 1.8 GByte Speicherkapazität erhältlich. Dieses Laufwerk . . . Das CD-ROM-Laufwerk wird von der Zeta-Maschinen GmbH geliefert, die ... Von derselben Firma wird auch ein Datenbanksystem für den Rechner vertrieben, das ..."

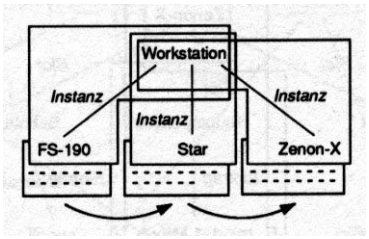
Als Kohärenzgraph:



- Das Progressionsmuster *abgeleiteter Themen* liegt vor, wenn in Folge verschiedene Konzepte (Themen) angesprochen werden, die alle einen gemeinsamen Oberbegriff besitzen. Im Beispiel:

  "Der FS-190-Rechner ... In bezug auf seine Verarbeitungsgeschwindigkeit steht der FS-190 dem Star von Star Inc. nicht nach, der .... Beide werden jedoch noch übertroffen von dem neuen Zenon-X. Er verfugt über ..."

  Als Kohärenzgraph:



Die obigen drei Progressionsmuster sind Grundtypen, die in verschiedenen Kombinationen auftreten können. Progressionsmuster eignen sich in besonderer Weise als *inhaltlich* motivierte Navigationspfade (vgl. Trigg 88 und Hahn 90) in Hypertexten, da sie die thematische Organisation eines Textes innerhalb von Absatzgrenzen sowie absatzübergreifend widerspiegeln. Die Kohärenzgraphen sind in den in Kapitel 2.1 beschriebenen Textgraph integriert, wobei die Textknoten in den Kohärenzgraphen den Textknoten im Textgraph entsprechen.

2.3 Konstruierte Kohärenzrelationen

Die in den Textgraphen und der Textwissensbasis enthaltene thematische und propositionale Information kann einerseits, wie oben beschrieben, genutzt werden, die thematische Struktur der Originaltexte zu explizieren und dem Benutzer als Navigationshilfe zur Verfügung zu stellen. Damit wird eine selektive Rezeption von Textfragmenten *eines* Textes auf einem vom Benutzer zu wählenden Abstraktionsniveau möglich. Dem steht ein konstruktiver Zugang zu Textinformation gegenüber, der basierend auf dem propositionalen Gehalt der Textwissensbasis einzelne Textfragmente in einen neuen Kontext stellt (Hammwöhner/Thiel

87). Eine Zusammenstellung von unterschiedlichen Originaltexten entstammenden Textfragmenten in sinnvoller Lesefolge ermöglicht es, in der Textmenge implizit gegebene Zusammenhänge offenzulegen und dabei gleichzeitig Redundanz zu vermeiden, indem Wiederholung schon präsentierter Inhalte durch Ausschluß von Texteinheiten verhindert wird.

Die Relationen zwischen den Textfragmenten sind dabei im Prinzip die gleichen, wie sie in linearen Texten auch auftreten, der Unterschied besteht allein in dem konstruktiven, Textgrenzen überschreitenden Gebrauch dieser Relationen. Thematische Progressionsmuster, wie sie in Abschnitt 2.2 beschrieben wurden, können dazu benutzt werden, die thematische Struktur von Hypertextpfaden zu planen, indem z.B. Textfragmente unter dem Gesichtspunkt zusammengestellt werden, daß sie eine nicht redundante, aber möglichst vollständige Beschreibung eines Objekts (Progression mit konstantem Thema) ergeben. Aus dem propositionalen Gehalt der Textwissensbasis lassen sich, in Anlehnung an die *Rhetorische Struktur-Theorie* (Mann/Thompson 88) oder die von Hobbs (Hobbs 85) definierten Kohärenzrelationen, darüber hinaus semantisch fundierte Hypertext-Kanten ableiten, die eine differenziertere Dialogplanung erlauben. Das Spektrum der verfügbaren Relationen ist dabei begrenzt durch die Tiefe der Textanalyse und die Ausdrucksmöglichkeiten der in der Textwissensbasis eingesetzten Frame-Repräsentationssprache, so daß zur Zeit weder temporal, noch kausal oder epistemisch begründete Relationen spezifiziert werden können. Es erweisen sich aber auch einfachere Kohärenzrelationen als genügend aussagekräftig. Im folgenden sind einige Beispiele aufgeführt (wobei die Textfragmente aus einem Text oder aus unterschiedlichen Texten stammen können):

- *Elaboration:* Ein in der ersten Texteinheit eingeführter Sachverhalt wird detaillierter dargestellt. "Der *Zenon-X* verfügt über eine leistungsstarke Graphiksoftware." — "Für den *Zenon-X* sind die Graphik-Pakete *Zen-Draw* und *Zen-Paint* verfügbar."
- *Bestätigung:* Ein Sachverhalt wird wiederholt behauptet oder auf allgemeineres zurückgeführt. "Der *Zenon-X* läuft unter *Zenix."* — "Für die Rechner der *Zenon-Baureihe* steht das Betriebssystem *Zenix* zur Verfügung."
- *Widerspruch:* Zwei Texteinheiten konstatieren unvereinbare Aussagen. "Der *Zenon-X* hat einen 50356-Prozessor." — "Die Rechner der *Zenon*-Baureihe haben einen *68030*-Prozessor."

3. Inhaltsorientierte Navigation und Präsentation von Textinformation

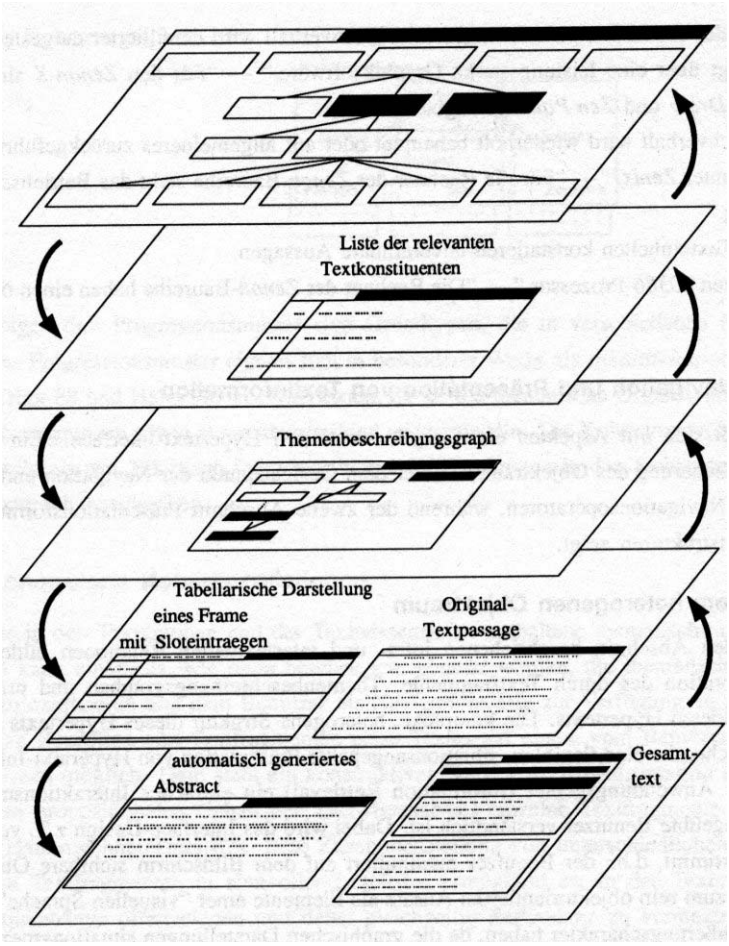
Dieses Kapitel befaßt sich mit Aspekten eines objektorientierten Hypertext-Interface. Ein erster Abschnitt erläutert die Gliederung des Objektraumes unter dem Gesichtspunkt der Navigation und führt zur Definition einfacher Navigationsoperatoren, während der zweite Abschnitt Präsentationsformen für die komplexen Hypertextstrukturen zeigt.

3.1 Navigation in einem heterogenen Objektraum

Die im vorangehenden Abschnitt beschriebenen intra- und intertextuellen Relationen bilden die Grundlage für die Exploration des durch Textfragmente, Themenbeschreibungsgraphen und propositionales Weltwissen gebildeten Hypertexts. Die komplexe, heterogene Struktur dieses Hypertexts erfordert eine Benutzeroberfläche, die eine flexiblen, situationsangepaßte Präsentation von Hypertext-Inhalten ermöglicht, während das Anwendungsgebiet (Information Retrieval) ein einfaches Interaktionsmodell erfordert, das auch für ungeübte Benutzer verständlich ist. Dabei wird das Interface-Design z.T. von der räumlichen Metapher bestimmt, d.h. der Benutzer manipuliert auf dem Bildschirm sichtbare Objekte, die jedoch im Gegensatz zum rein objektorientierten Ansatz als Elemente einer "visuellen Sprache" (Lakin

87, Thiel 89) auch Äußerungscharakter haben, da die graphischen Darstellungen situationsspezifisch erzeugt werden. Wichtig ist insbesondere auch eine Anpassung der Navigation an die Organisation des Objektraumes, der sich wie folgt darstellt:

- Für jedes Textfragment ist eine thematische und eine propositionale Repräsentation gegeben, nämlich ein Blattknoten des Textgraphen und die Textwissensbasis. Während die in den Knoten des Textgraphen enthaltenen semantischen Netze wegen ihres geringen Umfangs holistisch dargestellt und wahrgenommen werden können, bildet die frame-orientierte Textwissensbasis einen gesonderten Objektraum, der explorativ erkundet werden kann.
- Auf den thematischen Repräsentationen baut sich mit den Textgraphen eine Hierarchie thematischer Abstraktionen auf, die gleichzeitig als Zusammenfassungen und als Indexstruktur dienen können.
- Die propositionale Repräsentation ist die Grundlage für eine semantische Vernetzung von Texteinheiten durch Kohärenzrelationen, wie sie in Abschnitt 2.3 beschrieben wurden.



Relationierte Konzepte des Diskursgebiets aus der Wissensbasis

Liste der Passagensubstrate

Graphische Visualisierungen der thematischen Struktur von einzelnen Passagen

Faktische und textuelle Detailinformation

Abbildung 2 Stufen der kaskadierten Kondensierung in TWRM-TOPOGRAPHIC (Die Darstellung ist eine dem hier zugrundegelegten Implementierungsstand angeglichene Version der Illustration in Kuhlen et al. 89)

In dem von diesen Objekten aufgespannten Raum lassen sich zwei orthogonale Navigationsrichtungen unterscheiden:

- Navigation zwischen Textfragmenten bzw. den aus ihnen abgeleiteten Hypertext-Einheiten bei gleichbleibendem Abstraktionsniveau. Diese *horizontale* Navigation *(Browsing)* führt z.B. entlang einer durch ein thematisches Progressionsmuster motivierten Hypertext-Kante von der Themenbeschreibung eines Textfragmentes zu

der eines anderen, kann aber auch innerhalb des in sich geschlossenen Objektraums einer Textwissensbasis zwischen Frames erfolgen.

- Navigation zwischen Hypertexteinheiten unterschiedlichen Abstraktionsgrades. Diese *vertikale* Navigation *(Zooming,* Thiel/Hammwöhner 87) folgt z.B. den Abstraktionsrelationen im Textgraphen, verbindet dessen Blattknoten mit den zugehörigen Textfragmenten, bzw. verweist von Themen des Textgraphen auf Fakten im Textwissen, die diesen Themen zuzuordnen sind. Durch eine Folge von Zoom-Operationen läßt sich, ausgehend von genetischen Themenbeschreibungen, der Inhalt eines Textfragments im Sinne eines kaskadierten Abstracting (s. Abb. 2) sukzessiv erschließen.

Stehen Operatoren zur Verfügung, die *Browsing* und *Zooming* realisieren, ist von jedem beliebigen Teilobjekt der gesamte Hypertext explorativ zu erreichen. Ein geeigneter Startpunkt für die Navigation wird durch Abgleich der Themenstruktur von Texteinheiten mit einem Interessenprofil gefunden. Die Formulierung dieses Interessenprofils geschieht vor Beginn der Exploration durch Auswahl *(Selecting)* von Themen aus einer speziellen Wissensbasis, die das taxonomische Grundwissen des Diskursbereichs enthält, bzw. während der Navigation aus dem Textgraphen.

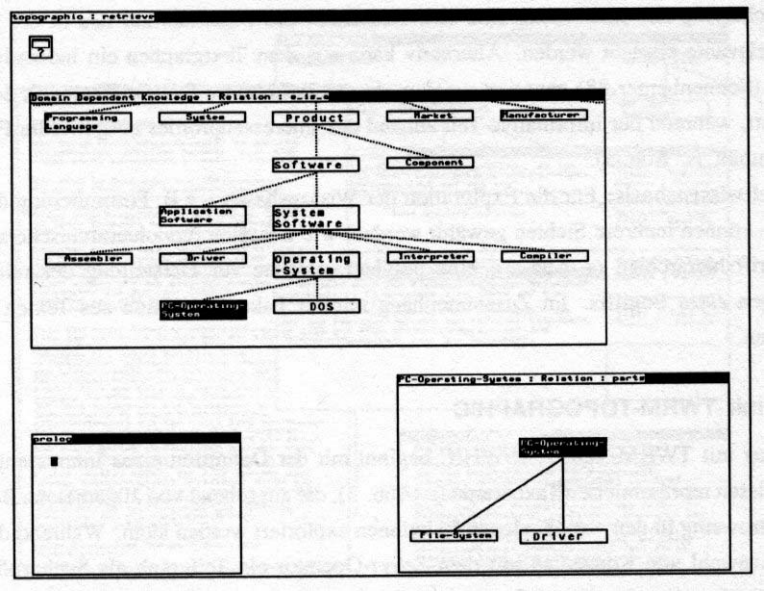


Abbildung 3 Monohierarchische Darstellung des Weltwissens in zwei Relationen, einer Spezialisierungs-, und einer Teil-von-Relation. Für das Suchprofil ausgewählte Begriffe sind invertiert dargestellt.

3.2 Präsentation von Textinformation

Neben einer adäquaten Strukturierung des Objektraumes ist die Auswahl der Präsentationsformen zentral für ergonomische Gestaltung einer Benutzerschnittstelle. Dabei können für die gleichen Informationsobjekte in unterschiedlichen Dialogsituationen verschiedene Darstellungen angemessen sein. Die bisher eingeführten Informationsobjekte Textfragment, Textgraph und Textwissen können in TWRM-TOPOGRAPHIC wie folgt aufbereitet werden:

- **Textfragment:** Der Fließtext als solcher ist kaum für situationsspezifische Darstellungsvarianten (von marginalen Änderungen, z.B. der Typographie etc. abgesehen) geeignet, aber die graphische Realisierung der Einbettung in den Kontext ist für die Rezeption wichtig. Bei der auf die Suche nach ergänzenden Fakten ausgerichteten intertextuellen Navigation wird eine Texteinheit als weitgehend eigenständiges isoliertes Objekt aufgefaßt, während bei Einbettung eines

Textfragments in die ursprüngliche Abfolge des Originaltextes eine die Textkohäsion unterstützende Darstellung angemessen ist (indem z.B. der Originaltext als ein geschlossenes Objekt dargestellt wird, wobei gleichzeitig die Interaktion weniger den Aspekt der Navigation als den des Blätterns hervorhebt - s. Abb. 5).

- **Textgraph:** Die semantischen Netze der Textgraphknoten erlauben in graphischer Darstellung eine holistische Wahrnehmung der Thematik einer Textpassage (bzw. eines Clusters von Textpassagen) und damit eine schnelle Einschätzung ihrer Relevanz. Im Kontext des gesamten, häufig sehr umfänglichen Textgraphen wird eine graphische Aufbereitung schnell unübersichtlich, nicht umsonst gibt Abb. l einen nur aus drei Passagen abgeleiteten Textgraphen wieder. In diesem Fall kann unter Berücksichtigung des Interessenprofils eine Reduktion des Materials auf das in der jeweiligen Situation relevante erreicht werden. Alternativ kann aus dem Textgraphen ein indikativ-informatives Abstract (Sonnenberger 88) generiert werden, dessen indikativer Teil die Thematik des Textes charakterisiert, während der informative Teil anhand des Interessenprofils ausgewählte Fakteninformationen enthält (s. Abb. 5).
- **Text-/Weltwissensbasis:** Für die Exploration der Wissensbasen - z.B. Formulierung des Interessenprofils - können mehrere Sichten gewählt werden, wie z.B. eine monohierarchische zur Erkundung von Begriffshierarchien (s. Abb. 3), eine polyhierarchische zur Darstellung der relationalen Verknüpfungen *eines* Begriffs. Im Zusammenhang mit der Faktenextraktion aus Texten sind Tabellen vorgesehen.

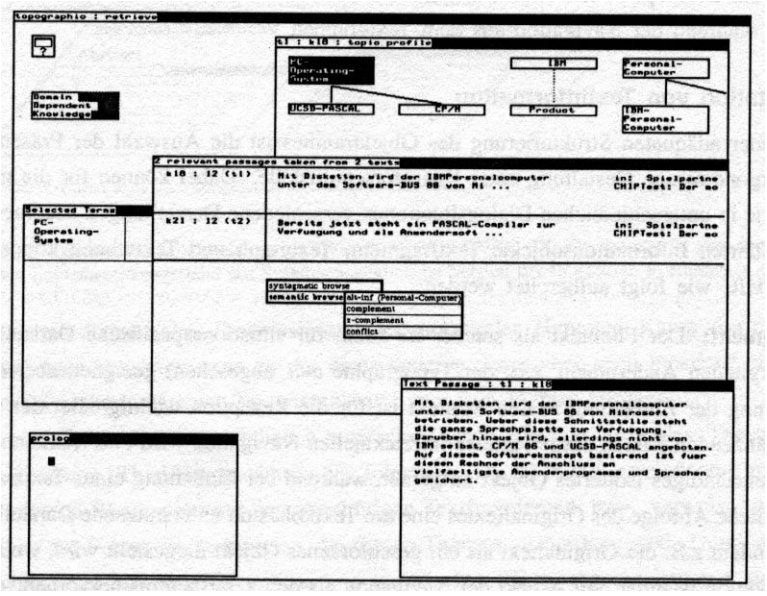


Abbildung 4 Der abgebildete Dialogzustand erklärt sich wie folgt: Ausgehend von der Liste der relevanten Passagen wurde das Themenprofil des relevantesten Textfragments erkundet, das zusätzlich in seiner textuellen Form präsentiert wurde. Im Augenblick wird durch Präsentation eines Menüs, das die von dem aktuellen Textfragment ausgehenden Hypertext-Kanten anführt, die Navigation zu einem semantisch verwandten Textfragment vorbereitet (dabei bezeichnen *complement* und *x-complement* elaborative Relationen, während *conflict* auf einen Widerspruch hindeutet. Mit einem Konzept zusätzlich indizierte Relationen, wie *alt-inf(Personal-Computer),* die einen Kontrast bezeichnet, ermöglichen die Vorgabe eines thematischen Fokus.

3.3 Dialog mit TWRM-TOPOGRAPHIC

Der Dialog mit TWRM-TOPOGRAPHIC beginnt mit der Definition eines Interessenprofils anhand der im Weltwissen repräsentierten Taxonomie (s. Abb. 3), die ausgehend von allgemeinen Begriffen durch sukzessives Browsing in den verschiedenen Relationen exploriert werden kann. Während der Navigation wird durch Auswahl von Konzepten mit dem *Select-Operator* ein Teilgraph als Suchprofil ausgewählt. Eine Übersicht über das

aktuelle Profil (als Tabelle und/oder Graph), das als thematische Beschreibung eines Clusters relevanter Textfragmente aufgefaßt werden kann, erhält der Benutzer durch *Zooming* auf die Gesamtdarstellung des Weltwissens. Ein weiteres *Zooming* stellt den Schritt von der thematischen Beschreibung zu einer Liste der relevanten Textfragmente dar, deren Einträge nach absteigender Relevanz sortiert sind (s. Abb. 4).

Diese durch partielles Matching (Hammwöhner/Thiel 87) gefundenen Textfragmente stellen den Ausgangspunkt für die Exploration des Hypertexts dar. Zunächst kann durch *Zooming* zu jeder Passage ein Themenprofil und, ausgehend von diesem, tabellarische Fakteninformation und der Volltext erreicht werden (s. Abb. 4 und 5). Orthogonal zu dieser Navigation in Richtung wachsender Spezifität kann auf jeder dieser Ebenen durch *Browsing* zu einer inhaltlich (oder syntaktisch) benachbarten Texteinheit navigiert werden. Die im aktuellen Kontext sinnvollen Relationen werden bei Bedarf in einem Menü angeboten (s. Abb. 4).

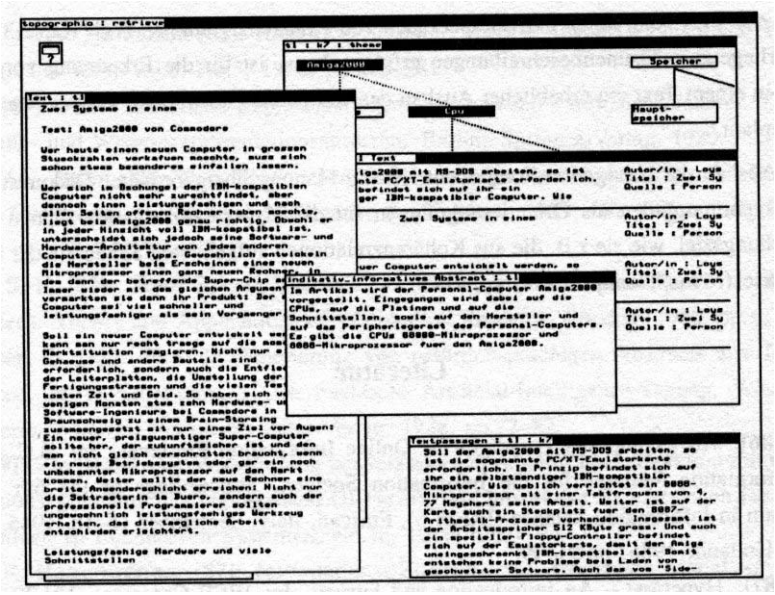


Abbildung 5 Die gezeigte fortgeschrittene Dialogsituation illustriert die Möglichkeiten zur gleichzeitigen Präsentation von lokaler und globaler Information, wobei sowohl textuelle als auch graphische Stilmittel eingesetzt werden: So geben der Themenbeschreibungsgraph (oben rechts) und die zugehörige Passage unterschiedlich detailliert die Inhalte des Textfragments mit dem (internen) Bezeichner k7 als lokale Information an, die ergänzt wird durch globale Kontextinformation in Form des situationsadäquat erzeugten Abstracts und des Volltextes (Mitte bzw. links). Darüber hinaus stehen weitere Textpassagen, die auch thematisch relevant sind, zur Auswahl (Mitte rechts). (Quelle: Kuhlen et al. 89)

4. Ausblick

Das wichtigste zukünftige Forschungsziel hinsichtlich der Textanalyse ist die Erweiterung der von TOPIC/TOPOGRAPHIC unterstützten Hypertextgraphen um die Berücksichtigung von Kohärenzrelationen, die nicht auf die in Kapitel 2.2 beschriebenen thematischen Progressionsmuster zurückgehen, sondern *Argumentationsmuster* in einem Text anzeigen (vgl. z.B. Mann/Thompson 88). Beispiele hierfür sind die Gegenüberstellung verschiedener Aussagen zu einem Sachverhalt, die zeitlichen Beziehungen zwischen verschiedenen, in einem Text beschriebenen Ereignissen oder die Begründung einer zuvor aufgestellten Behauptung. Die Schwierigkeit mit der Unterstützung solcher Kohärenzrelationen liegt in dem enormen Aufwand, der nötig ist, um sie automatisch (wie die anderen Relationen auch) aus einem Text abzuleiten. Anders als beim konstruktiven Gebrauch von Kohärenzrelationen (vgl. Kap.2.3), der auf der Basis der vorliegenden Themenbeschreibungen erfolgen kann, ist

für die Erkennung von Argumentationsmustern in einem Text ein erheblicher Ausbau des Textparsers notwendig. Eine solche Erweiterung ist jedoch geplant.

Für den Ausbau der Navigationskomponente ist die Planung übergeordneter Diskursstrukturen, die analog den Hypertextpfaden als Orientierungsrahmen für die Hypertextnavigation dienen können, das nächste Forschungsziel, wie sie z.B. die aus Kohärenzrelationen aufgebauten Schemata der Rhetorischen Struktur-Theorie (Mann/Thompson 88) vorsehen.

Literatur

Bates, M.J. [86]: An Exploratory Paradigm for Online Information Retrieval. In: B.C. Brooks (ed): Intelligent Information Systems for the Information Society. Proceedings of the 6th International Research Forum in Information Science (IRFIS 6), Frascati, Italy, September 16-18, 1985. Amsterdam et al: North Holland, 1986, pp. 91-99.

Conklin, J. [87]: Hypertext - An Introduction and Survey. In: IEEE Computer, Vol. 20, No. 9, 1987, pp. 17-41.

Daneš, F. [74]: Functional Sentence Perspective and the Organization of the Text. In: F. DaneS (ed): Papers on Functional Sentence Perspective. Prague: Academia, 1974, pp. 106-128.

Frisse, M.E. [88]: From Text to Hypertext. In: Byte, Vol. 13, No.10, 1988, pp. 247-253.

Hahn, U. [90]: Topic Parsing: Accounting for Text Macro Structures in Full-Text Analysis. In: Information Processing & Management 26. 1990, No. l.

Hahn, U. / U. Reimer [86]: Semantic Parsing and Summarizing of Technical Texts in the TOPIC System. In: R. Kuhlen (ed): Informationslinguistik. Tübingen: Niemeyer, 1986, pp. 153-193.

Hahn, U. / U. Reimer [88]: Automatic Generation of Hypertext Knowledge Bases. In: Proc. of the Conf. on Office Information Systems, 1988. New York: ACM Press, 1988, pp. 182-188.

Halliday, M.A.K. / R. Hasan [76]: Cohesion in English. London: Longman, 1976.

Hammwöhner, R. / U. Thiel [87]: Content Oriented Relations between Text Units - A Structural Model for Hypertexts. In: Hypertext '87 Papers, Chapel Hill, NC, University of North Carolina, 1987, pp. 155-174.

Hobbs, J.R. [85]: On the Coherence and Structure of Discourse. Stanford University, Report CSLI-85-37, 1985.

Kuhlen, R. / R. Hammwöhner / G. Sonnenberger / U. Thiel [89]: TWRM-TOPOGRAPHIC: Ein wissensbasiertes System zur situationsgerechten Aufbereitung und Präsentation von Textinformation in graphischen Retrievaldialogen. In: Informatik Forschung und Entwicklung, Vol. 4, No. 2, 1989, pp. 89-107.

Lakin, F. [87]: Visual Grammars for Visual Languages. In: Proc. 6th National Conf. on Artificial Intelligence, 1987, pp. 683-688.

Lenat, D.B. / A. Borning / D. McDonald / C. Taylor / S. Weyer [83]: Knoesphere - Building Expert Systems with Encyclopedic Knowledge. In: Proc. 8th Int. Joint Conf. on Artificial Intelligence, 1983, pp. 167-169.

Mann, W.C. / S.A. Thompson [88]: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. In: Text, Vol. 8, No. 3, 1988, pp. 243-281.

Reimer, U. [89]: FRM: Ein Frame-Repräsentationsmodell und seine formale Semantik. Zur Integration von Datenbank- und Wissensrepräsentationsansätzen. Berlin: Springer-Verlag, 1989.

Reimer, U. / U. Hahn [88]: Text Condensation as Knowledge Base Abstraction. In: Proc. of the 4th IEEE/AAAI Conference on Artificial Intelligence Applications, 1988. Washington: Computer Society Press, 1988, pp. 338-344.

Robertson, S. E. [80]: Some Recent Theories and Models in Information Retrieval. In: O. Harbo, C. Kajberg (ed): Theory and Applications of Information Research, London, 1980, pp. 131-136.

Sonnenberger, G. [88]: Flexible Generierung von natürlichsprachigen Abstracts aus Textrepräsentationsstrukturen. In: H. Trost (ed): 4. Österreichische Artificial-Intelligence-Tagung. Wiener Workshop Wissensbasierte Systeme. Berlin: Springer-Verlag, 1988, pp. 72-82.

Thiel, U. [89]: Zur illokutiven Modellierung konversationaler graphischer Interaktion mit wissensbasierten Informationssystemen. In: Tagungsband GI-Fachtagung Interaktive Schnittstellen für Informationssysteme, Notizen zu Interaktiven Systemen, No. 18, 1989, pp. 61-77.

Thiel, U. / R. Hammwöhner [87]: Informational Zooming: An Interaction Model for the Graphical Access to Text Knowledge Bases. In: C.T. Yu, CJ. van Rijsbergen (eds): Proceedings of the lOth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, 1987, pp. 45-56.

Thiel, U. / Hammwöhner, R. [89]: Interaktion mit Textwissensbasen: Ein objektorientierter Ansatz. In: Paul, M. (ed): Proc. GI- 19. Jahrestagung I, Berlin, Heidelberg, 1989, pp. 81-95.

Tiamiyu, M. / I.Y. Ajiferuke [88]: A Total Relevance and Document Interaction Effects Model for the Evaluation of Information Retrieval Processes. In: Information Processing & Management, Vol. 24, No. 4, 1988, pp.391-404.

Trigg, R.H. [88]: Guided Tours and Tabletops - Tools for Communicating in a Hypertext Environment. In: ACM Transactions on Office Information Systems, Vol. 6, No. 4, 1988, pp. 398-414.

Trigg, R.H. / M. Weiser [86]: TEXTNET: A Network-Based Approach to Text Handling. In: ACM Transactions on Office Information Systems, Vol.4, No.l, 1986, pp.1-23.

Weyer, S.A. / A.H. Borning [85]: A Prototype Electronic Encyclopedia. In: ACM Transactions on Office Information Systems, Vol. 3, No. l, 1985, pp. 63-88.