3.1 Allgemeine Angaben zum Teilprojekt D1

3.1.1 Thema:

Verwaltung des Datenkorpus für die SFB-Teilprojekte; korpusbasierte Implementierung eines Textproduktionsmodells mit dem Schwerpunkt Informationsstruktur

3.1.2 Fachgebiete und Arbeitsrichtung:

Computerlinguistik; Korpuslinguistik und Automatische Textgenerierung

3.1.3 Leiter:

Stede, Manfred Prof. Dr.
geb. am 05.02.1965

Universität Potsdam
Institut für Linguistik
Postfach 60 15 53
14415 Potsdam

Telefon: 0331 – 977-2691
Telefax: 0331 – 977-2761
E-Mail: stede@ling.uni-potsdam.de

3.2 Zusammenfassung

Das Projekt stellt dem SFB eine Infrastruktur zur Verwaltung und zur Abfrage der in den Teilprojekten erhobenen linguistischen Daten zur Verfügung (Service). Darüber hinaus verfolgt es Forschungsziele hinsichtlich der Repräsentation von Informationsstruktur in linguistischen Korpora und insbesondere zur Rolle der Informationsstruktur in Texten; diese Fragen sollen anhand eines Modell der automatischen Textproduktion bearbeitet werden.

Der Service-Anteil besteht darin, die Hardware- und Software-Voraussetzungen für eine zentrale Korpusverwaltung zu schaffen und allen Teilprojekten die Arbeit mit den sukzessive entstehenden Korpora über eine einfache WWW-Schnittstelle zu ermöglichen. Diese Schnittstellen sollen von vornherein so gestaltet sein, dass die Korpora auch für andere Forschungszwecke außerhalb des SFB genutzt werden können. Während die speziellen Erfordernisse für die Bearbeitung gesprochener Sprache im Projekt D3 berücksichtigt werden, konzentriert D1 sich auf geschriebene Daten. Da bisherige linguistische Datenbanken sich mit der Annotation von Informationsstruktur nur in ersten Ansätzen beschäftigt haben, stellt sich dabei das Forschungsziel, die systematische Integration informationsstruktureller Kategorien in bestehende Annotationsschemata (part-of-speech tags, syntaktische Strukturen) zu untersuchen und Lösungsvorschläge zu entwickeln. Perspektivisch ergibt sich weiterhin das Ziel, eine Abfragesprache zu entwickeln und zu implementieren, die einerseits die Kombination von IS- und anderen Suchkriterien erlaubt und andererseits ein intuitiv einfaches Recherchieren ermöglicht, das keine profunden Kenntnisse in formalen Sprachen voraussetzt, sondern durch intelligente Werkzeuge unterstützt wird.

Neben der Korpusarbeit besteht der Forschungsschwerpunkt des Projekts darin, die Rolle der Informationsstruktur für die Beschreibungsebene „Text“ zu untersuchen und sie in Modelle der Textstruktur zu integrieren. Beschränkt auf die Textsorte Zeitungskommentar, also auf argumentative Texte, sollen insbesondere die Zusammenhänge zwischen „rhetorischer Struktur“ des Textes und informationsstrukturellen Kategorien geklärt werden: Was ist ein geeignetes Beschreibungsinventar für die „Informationsstruktur von Texten“ und wie schlägt sie sich jeweils in der Informationsstruktur einzelner Sätze nieder? Diese Arbeiten sollen in ein formales Modell der Textproduktion münden, das – auf der Grundlage bereits existierender Komponenten – in einem automatischen Textgenerator implementiert wird. Gründlicher als in bisherigen Textgenerierungssystemen soll es möglich sein, bei gleich bleibender rhetorischer Struktur verschiedene Text-Varianten zu erzeugen, die hinsichtlich der Informationsstruktur variieren.

3.3 Stand der Forschung

Linguistische Datenbanken, Annotation und Abfrage

Im Zuge des Trends zur korpus-basierten Arbeit entstanden in der Computerlinguistik seit den späten 1980er Jahren eine Reihe von Datenbanken mit annotierten Sprachdaten, die mit part-of-speech tags oder mit (mitunter partiellen) syntaktischen Strukturen versehen sind. Als Klassiker gilt dabei die „Penn TreeBank“ (Marcus, Santorini & Marcinkiewicz 1993). Der eine Zweck solcher Datenbanken ist das Training statistischer oder probabilistischer Modelle, die dann zur automatischen Analyse ungesehener Daten verwendet werden; diese Ziele sind für unser Anliegen weniger relevant. Wichtig ist für uns hingegen die Nutzung von Datenmengen zur Orientierung von Linguisten bei der (manuellen) Analyse sowie für

statistische Auswertungen bestimmter Korrelationen in den gesammelten Daten. In dieser Hinsicht war „LDB“ (van Halteren & van den Heuvel 1990) ein frühes Beispiel für den Vorschlag eines Datenbanksystems, das Werkzeuge für die Recherche in syntaktisch annotierten Daten bereitstellt. Durch Angabe eines Baum-Gerüsts mit attributierten Knoten können alle „matchenden“ Strukturen aus der Datenbank abgerufen werden.

Ein aktuelleres Beispiel ist das an der Universität Saarbrücken erstellte „NEGRA“ Korpus (Skut et al. 1997). Es bietet 20.000 Sätze aus der „Frankfurter Rundschau“, die mit part-of-speech tags, grammatischen Funktionen und phrasalen Knoten annotiert sind. Technisch liegt NEGRA eine SQL-Datenbank zugrunde, und eine Reihe verschiedener Export-Formate (u.a. Penn TreeBank) dienen als Schnittstelle zu Software-Modulen oder zum menschlichen Betrachter. Das Werkzeug „ANNOTATE“ erlaubt die Erstellung syntaktischer Repräsentationen direkt am Bildschirm durch Visualisierung der entstehenden Strukturen. Durch die Einbindung von Taggern und Parsern kann die Strukturzuweisung halbautomatisch verlaufen: Die Software schlägt eine partielle Analyse vor, die von der NutzerIn bestätigt oder revidiert wird. Auf diesem Wege wird derzeit im „TIGER“ Projekt (Brants et al. 2002), an dem auch das Institut für Germanistik der Universität Potsdam beteiligt ist, das Korpus noch beträchtlich erweitert, zudem wird die Annotation auch auf morphologische Information ausgedehnt. Das gewählte Baumformat „combines the advantages of dependency grammar and phrase structure grammar“ (Brants et al. 2002); die Strukturen werden relativ flach gehalten, indem beispielsweise die Unterscheidung zwischen Argumenten und Adjunkten nicht in der Konstituentenstruktur ausgedrückt wird, sondern durch die am Knoten annotierte syntaktische Funktion. An den Blättern der Bäume wird das „Stuttgart-Tübingen Tagset“ (Schiller, Teufel & Stöckert 1999) für part-of-speech tags benutzt. Die für die Datenbank entwickelte Abfragesprache TIGERSearch (Lezius 2002) erlaubt die Suche auf drei verschiedenen Ebenen:

- Knoten: Boole’sche Ausdrücke über Attribut/Wert Paaren
- Relationen zwischen Knoten: lineare Präzedenz, Dominanz
- Graph-Beschreibungen: Eingeschränkte Boole’sche Ausdrücke über Relationen zwischen Knoten

Ganz ähnlich zu TIGERSearch gestattet auch VIQTORYA (Steiner & Kallmeyer 2002) das Recherchieren in Baumbanken; einige Unterschiede bestehen im Hinblick auf die in den zugrunde liegenden Korpora jeweils gewählten Kodierungsformate. Zur Illustration ein Beispiel aus (Steiner & Kallmeyer 2002): Der folgende Ausdruck sucht im Korpus nach der Präposition *von* gefolgt von der Präposition *bis*, mit der Zusatzbedingung, dass beide gemeinsam von einer Präpositionalphrase dominiert werden:

- token (1) = von & token (2) = bis & 1..2 & cat (3) = PX & 3 >>1 & 3 >> 2

Der an der Universität Tübingen entwickelte TUSNELDA Annotationsstandard (Wagner & Kallmeyer 2001) erfüllt für den SFB 441 etwa den gleichen Zweck wie das hier beantragte Teilprojekt für den hier beantragten SFB. TUSNELDA bündelt die in den verschiedenen Teilprojekten gesammelten Daten und entwickelt ein Annotationsformat, das die unterschiedlichen Bedürfnisse der Projekte gleichermaßen abdeckt. Das Korpusformat orientiert sich an der XML-Version des ursprünglich in SGML formulierten Corpus Encoding Standard, XCES (siehe http://www.cs.vassar.edu/XCES), und die Abfragesprache ist eng an die an der Edinburgh University entwickelte „XML QUERY“ angelehnt. Es wurden einige Erweiterungen dieser Standards vorgenommen, u.a. um auch gesprochene Dialoge und Comics bearbeiten zu können. Auch hier sei ein Beispiel für eine Suchanfrage genannt: Der

folgende Term „matches every <figure> element which contains a situational characterization with the keywords „forefinger“ and „bent“.

- /figure!/.*/situation/keywords/(term/#“forefinger“ & term/#bent)

Festzuhalten ist an dieser Stelle, dass eine Reihe syntaktisch orientierter Datenbanken existieren, die für Forschungszwecke verfügbar sind. Die Annotation von informationsstrukturellen Kategorien ist in solchen Datenbanken unseres Wissens bislang nicht realisiert worden; gegenwärtig beginnt an der Universität des Saarlandes ein Projekt (MULI), das sich dieser Problematik widmet, und mit dem wir kooperieren wollen. Unser Projekt soll auf den Arbeiten von NEGRA/TIGER und TUSNELDA aufbauen und Datenformate definieren, die auf die Bedürfnisse der Teilprojekte des SFB abgestimmt sind. Ein Schwerpunkt wird dabei die Integration von Informationsstruktur-Annotation in syntaktische Baum-Repräsentationen sein. Zur Abfrage der entstehenden Datenbanken soll perspektivisch eine *query language* entworfen und implementiert werden, die benutzerfreundlicher ist als die oben gezeigten Beispiele aus formalen Abfragesprachen.

Textstruktur

In der Tradition der von Grimes (1975) vorgeschlagenen *rhetorical predicates* entstanden in den 1980er Jahren eine Reihe von Ansätzen zur Beschreibung der Struktur von Texten auf der Basis von Diskursrelationen, die zwischen benachbarten Textteilen bestehen. Die Vorschläge unterschieden sich hinsichtlich der Disziplin, aus der sie stammten und setzten dementsprechend die Akzente unterschiedlich. So betonte etwa Hobbs (1985), beeinflusst von der Artificial Intelligence, den Zusammenhang zwischen Welt-Wissen und der Identifikation von Diskursrelationen im Text. Primär aus der Psycholinguistik motiviert war demgegenüber der Vorschlag von Sanders, Spooren & Noordman (1992), die experimentell Evidenz für bestimmte Relationen zu gewinnen suchten und darüber hinaus versuchten, die Definitionen von Diskursrelationen kompositional aus elementaren Bausteinen anzugeben.

Einen anderen wichtigen Schritt unternahmen Mann & Thompson (1988) mit ihrer *Rhetorical Structure Theory* (RST), indem sie nicht allein eine Menge von etwa 20 Relationen vorschlugen, die für die Beschreibung nahezu beliebiger Texte geeignet sein sollen, sondern auch Kriterien für die Wohlgeformtheit der Textstruktur formulierten, die durch rekursive Anwendung der Relationen auf immer größere Texteinheiten entsteht. Mann & Thompson versuchten mit RST den Begriff der Kohärenz zu explizieren, was sich nicht nur in der Textlinguistik, sondern auch in der Automatischen Textgenerierung (s.u.) als sehr einflussreich erwies, zumal die Relationen – informell, doch systematisch – durch Angabe der jeweils verfolgten kommunikativen Ziele definiert sind, was eine Verwendung in Textproduktionsmodellen nahelegt. Im Gegensatz zur empirisch motivierten RST ging Asher (1993) das Problem aus der Perspektive der formalen Semantik an und schlug für eine kleine Menge von Relationen präzise Definitionen auf der Grundlage einer Default-Logik vor.

Hinsichtlich der Zahl der Relationen erheblich „sparsamer“ war der Vorschlag von Grosz & Sidner (1986), wonach zwischen benachbarten Textsegmenten – genauer: den ihnen zugrunde liegenden Zielen oder „discourse purposes“ – entweder eine „satisfaction-precedence“ oder eine „dominance“ Beziehung besteht: Entweder geht Ziel A dem Ziel B voran, oder Ziel A ist Unter-Ziel von Ziel B. Ganz ähnlich war der Vorschlag von Brandt & Rosengren (1992) im Rahmen des „Sprache und Pragmatik“-Projekts, die allerdings die „stützende“ Funktion von Textsegmenten/Illokutionen noch feiner gliederten in „subsidiäre“ und „komplementäre“

Funktion; erstere tragen direkt, letztere nur indirekt zum Erfolg der übergeordneten Illokution bei.

Die genannten Ansätze sind durchweg als „oberflächenfern" einzustufen, sie machen keine oder nur sehr vage Aussagen darüber, wie sich eine bestimmte Relation an der Textoberfläche niederschlägt, etwa durch Konnektoren oder Interpunktionszeichen. Demgegenüber bietet Martin (1992) für das Englische seinen Vorschlag der *conjunctive relations* (CR), die von den Merkmalen der linguistischen Realisierung ausgehen und über diese eine semantische Klassifikation in inhaltliche Relationen legen. CR ist somit eher ein Ansatz zur Beschreibung von *clause*-Verknüpfungsmöglichkeiten als eine Theorie der Textstruktur.

Semantische und pragmatische Relationen sind eine wichtige Beschreibungsebene für die Struktur von Texten, jedoch nicht die einzige. Von Bedeutung sind für unser Interesse hier vor allem noch Vorschläge zum Umgang mit „referentieller Struktur" sowie „thematischer Struktur" bzw. Informationsstruktur. Referentielle Struktur betrifft aus Analyse-Sicht die Auflösung anaphorischer Bezüge, aus Produktions-Sicht bei der Bezeichnung eines Diskursreferenten die Auswahl aus dem Spektrum der von der Sprache angebotenen referierenden Ausdrücke; sie hängt davon ab, ob/wie auf den fraglichen Diskursreferenten zuvor im Text Bezug genommen wurde, und welche alternativen Diskursreferenten von einem bestimmten referierenden Ausdruck fälschlicherweise identifiziert werden könnten. Einflussreich waren hier eine ganze Reihe von Arbeiten, die auf den „Centering" Ansatz nach Grosz, Joshi & Weinstein (1995) zurückgehen. Centering beinhaltet eine diskrete und relativ informationsarme Repräsentation, die einem Diskurs-Segment je einen einzelnen Diskursreferenten als „backward looking center" sowie eine geordnete Menge von Referenten als „forward looking center" zuordnen. Für die Bestimmung der Ordnung wurden verschiedene Algorithmen vorgeschlagen, wobei man dann auch soweit gehen kann, die diskrete Repräsentation zu verlassen und für jeden Diskursreferenten an jedem Punkt des Diskurses seinen jeweiligen numerischen „Aktivierungsgrad" zu errechnen, der dann entscheidend beeinflusst, mit welchen sprachlichen Mitteln auf ihn referiert werden sollte. Stellvertretend für diese aus der Psycholinguistik stammenden Ansätze sei die „accessibility theory" von Ariel (2001) genannt.

Kann nun der Begriff der Informationsstruktur auch für die Beschreibungsebene „Text" sinnvoll verwendet werden? Wenn ja, was soll er ausdrücken? Im o.g. „Sprache und Pragmatik" Projekt wurde der Terminus „kommunikative Gewichtung" verwendet (Brandt 1996): Gesteuert von den Intentionen der Textproduzentin werden bestimmte Textelemente eher in den Vorder- oder in den Hintergrund gerückt. Brandt spricht von der globalen Informationsstruktur als Gegenstück zur lokalen Informationsstruktur im Satz und untersucht die Korrelation zwischen der kommunikativen Gewichtung und den Strukturen komplexer Sätze im Text. So sei Subordination ein Ausdruck kommunikativer Gewichtung, sofern in der Tat die Wahl bestand, Subordination zu realisieren oder nicht. Eines ihrer Beispiele:

(1) Häberle wohnt in Tübingen und arbeitet in Stuttgart.

(2) Häberle, der in Tübingen wohnt, arbeitet in Stuttgart.

Während (1) eine einzelne Informationseinheit übermittele, trage (2) zwei getrennte Informationen, die unterschiedlich gewichtet seien.

Eine andere Perspektive nehmen funktional orientierte Ansätze ein, die im Text „thematische Entwicklung" untersuchen. So demonstrierte etwa Fries (1981), wie die Wahl der Satzgliedstellung die „connectivity" in einem Text beeinflusst und legt dabei (wie Halliday) für das Englische einen positionalen Begriff von Theme (bis zum finiten Verb) und Rheme (nach dem finiten Verb) zugrunde. Zu unterscheiden sind hier textsortenunabhängige Faktoren, die

den Zusammenhalt des Texts betreffen und textsortenspezifische Konventionen für die Wahl von Satz-themes, wie sie etwa von Ramm & Villiger (1995) für die Textsorte „Reiseführer" untersucht wurden. Der Gedanke hier ist, dass einer Textsorte eine bestimmte „chaining strategy" (siehe dazu auch Lavid 2000) zugrunde liegt, die als Satz-Thema stets Konstituenten eines bestimmten semantischen Typs bevorzugt; bei den genannten Reiseführern sind es Lokation-Konzepte.

Abschließend seien kurz einige Arbeiten genannt, die die Wechselwirkungen zwischen den genannten Beschreibungsebenen untersuchen. Grosz & Sidner (1986) hatten darauf hingewiesen, dass die intentionale Struktur eine Segmentierung des Diskurses liefert, die Konsequenzen für die referentiellen Ausdrücke, vor allem für die Pronominalisierung hat. Ähnlich haben sich Klein & von Stutterheim (1992) mit den Zusammenhängen zwischen Textstruktur und Referenz aus Sicht des „quaestio"-Ansatzes beschäftigt. Für die RST machten Knott et al. (2001) die Beobachtung, dass die Diskursrelation *Elaboration* (für Objekte/Attribute) einen anderen Status besitze als die übrigen „rhetorischen Relationen", da sie stets mit „focus shift" (in der Terminologie des SFB wohl „Topik-Bewegung") einhergehe; sie solle daher aus dem Relationsinventar entfernt und die entsprechenden RST-Teilbäume durch eine semantisch leere Topikwechsel-Relation verbunden werden. Unser Projekt möchte diesen Gedanken aufnehmen, sieht aber das von Knott et al. (2001) aufgeworfene Problem nicht als Eigenschaft der Elaborationsrelation, sondern vielmehr als Hinweis auf die allgemeine Situation, dass rhetorische Struktur und Topikwechsel sich zwar beeinflussen, jedoch separate Beschreibungsebenen sind, die auch als solche zu behandeln und nicht in einer einzigen Baumstruktur unterzubringen sind.

Automatische Textgenerierung

In der Automatischen Textgenerierung hat sich in den letzten Jahren das u.a. in (Reiter & Dale 2000) beschriebene „Pipeline" Modell als Standard etabliert, das den Textproduktionsprozess als dreistufig charakterisiert: In der *Textplanung* wird zunächst aus dem kommunikativen Ziel und den zu übermittelnden Informationen ein Textplan erstellt, der in der Regel als Baumstruktur aus Propositionen (an den Blättern) und Diskursrelationen (an den internen Knoten) aufgefasst wird. Die *Satzplanung* linearisiert diesen Baum zu einer Folge von satzsemantischen Spezifikationen, die in der Regel auch bereits lexikalisiert sind und die Satzstruktur festlegen. Durch Verwendung grammatischen Wissens über die Zielsprache bestimmt die *Oberflächenrealisierung* daraus schließlich wohlgeformte Sätze. Für unser Vorhaben sind zunächst besonders die Textplanung und die Satzplanung von Interesse (später auch die Realisierung). Viele Systeme legen der Textplanung die o.g. RST zugrunde (z.B. Moore & Paris 1993; O'Donnell et al. 2001), indem sie die Diskursrelationen als Planungsoperatoren formulieren und dann entweder nach einem top-down oder einem bottom-up Planungsalgorithmus aus den Eingabe-Daten einen wohlgeformten RST-Baum berechnen. Die vorgeschlagenen Systeme gehen davon aus, dass im Textplan die Abfolge der zu kommunizierenden Informationseinheiten bereits festgelegt ist, was allerdings eine starke Vereinfachung darstellt und die Berücksichtigung von informationsstrukturellen Faktoren bei der Textproduktion erheblich erschwert (vgl. 3.5).

Die Satzplanung ist erst seit wenigen Jahren als wichtige, eigenständige Komponente der Textgenerierung erkannt worden und ist dementsprechend noch nicht so ausführlich untersucht wie Textplanung und Oberflächenrealisierung, die heute mit fertigen Software-Komponenten wie z.B. KPML (Bateman 1997) gelöst werden kann. Da die verschiedenen Aufgaben der

Satzplanung (Wortwahl, referierende Ausdrücke, Konnektoren, Satzstruktur) vielfältig miteinander interagieren, ist eine strikte Sequenzialisierung nur auf Kosten der „Ausdrucksstärke“ des Generators möglich. Daher haben Wanner & Hovy (1996) ein flexibles Blackboard-basiertes Verfahren vorgeschlagen, das jedoch nicht vollständig implementiert wurde. Stone & Doran (1997) betonen hingegen die enge Verbindung zwischen Satzplanung und Realisierung und beziehen bereits Teile der zielsprachlichen Grammatik mit ein.

Für bestimmte Textsorten wie z.B. technische Anleitungen oder medizinische Berichte hat sich das Pipeline-Modell in der Praxis gut bewährt, und unser Projekt wird im Prinzip auch daran festhalten, jedoch die Aufgabenteilung zwischen Text- und Satzplanung neu gestalten und dazu ein Diskursgedächtnis als wesentliche Informationsquelle einsetzen. Bisherige Generatoren verwenden lediglich Listen von bislang verwendeten referierenden Ausdrücken – dies genügt für die Wahl des jeweils nächsten referierenden Ausdrucks im Text, nicht jedoch für weitere Entscheidungen über die lineare Abfolge von Konstituenten unter informationsstrukturellen Gesichtspunkten.

3.4 Eigene Vorarbeiten

Die Computerlinguistik-Arbeitsgruppe des Instituts hat bislang keinen korpuslinguistischen Forschungsschwerpunkt, setzt jedoch eine Reihe von Textkorpora für verschiedene Zwecke ein. Zudem wurde als Eigenentwicklung für die Arbeit mit Textstrukturen in einem studentischen Projekt ein eigenes Korpus mit 120 Zeitungskommentaren zusammengestellt und mit dem „RST-Tool“ (O’Donnell 1997) hinsichtlich der Diskursrelationen annotiert. (Abgesehen von den von Carlson, Marcu & Okurowski (2001) erstellten englischen Daten gibt es ansonsten noch keine verfügbaren Ressourcen dieser Art.) Dabei wurde auch ein eigenes XML-Format für diese Daten entwickelt und, darauf aufbauend, ein Recherche-Werkzeug implementiert, mit dem in dem Korpus nach Vorkommen bestimmter Diskursrelationen gesucht werden kann. Diese Ressourcen bilden einerseits eine Grundlage für die geplanten korpusbasierten Arbeiten zur Textstruktur; andererseits konnten dabei allgemeine Erfahrungen für die Verwaltung von Textkorpora und die Programmierung von Analyse-Werkzeugen unter Verwendung von XML/PERL gesammelt werden.

Manfred Stede hat sich seit 10 Jahren mit Textstruktur und vor allem mit Relations-basierten Ansätzen in der Tradition von Mann & Thompson (1988) beschäftigt (s. z.B. Rösner & Stede 1993; Grote, Lenke & Stede 1997). Diese Arbeiten standen stets im Kontext der Automatischen Textgenerierung. Eine erste auf RST basierende Implementierung eines Generators für Instruktionstexte war „Techdoc“ (Rösner & Stede 1992), auf dessen Grundlage dann zunächst die Rolle der Lexikalisierung für den Generierungsprozess ausführlich untersucht wurde (Stede 1999a). Die dabei entstandene Software-Komponente zur Lexikon-gesteuerten Überführung konzeptueller Repräsentationen in satzsemantische Spezifikationen kann nach Anpassungen für das beantragte Projekt verwendet werden. Einen zweiten Schwerpunkt bildete die Ausgestaltung der Textrepräsentation und dabei vor allem die korpusbasierte Untersuchung von Konnektoren als linguistische Signale für Diskursrelationen (z.B. Stede 1995). Das dabei entstandene „Diskursmarker-Lexikon“ (Grote & Stede 1998) wird derzeit für die Verwendung in einem Generator für Buchbeschreibungen und –empfehlungen aufbereitet (Stede 2002), in dem bereits ein prototypisches Textplanungsmodul entstanden ist. Auch diese Ressourcen stehen für das beantragte Projekt zur Verfügung. In (Stede 1999b) wird auf das Problem der Interaktion zwischen rhetorischer Struktur und

Informationsstruktur im Text hingewiesen, das bisherige Textgenerierungsmodelle nicht hinreichend berücksichtigt haben, und das im hier vorgeschlagenen Projekt grundlegend behandelt werden soll.

Michael Grabski hat an einer Modellierung des Zusammenhangs gearbeitet, der zwischen einem lokal in Texten definierten Diskurstopik, wie es in SDRT definiert ist (Asher 1993), und dem in der IS definierten Satztopik besteht. Für die Darstellung des Zusammenhangs hat sich als nützlich erwiesen, Texte zu analysieren, die nach der Diskursrelation ELABORATION strukturiert sind, und dabei die semantischen Eigenschaften der Diskursrelation selbst zu charakterisieren (Grabski 2000, 2001). Bestimmte satzinterne Markierungen der IS, u.a. die Linksversetzung, lassen sich damit als Resultate einer intendierten rhetorischen Struktur von Texten darstellen.

Christian Chiarcos hat in seiner Diplomarbeit (Chiarcos 2002) untersucht, welche Rolle die Salienz von Diskursreferenten für die Linearisierungsentscheidungen der Satzplanung spielt. Seine Konzeption und prototypische Implementierung können im Projekt unmittelbar weiter entwickelt werden. Als Diplom-Informatiker hat er zudem Erfahrungen mit dem Einsatz von Datenbanken und WWW-Servern.

Michael Götze beschäftigt sich gegenwärtig in seiner Diplomarbeit mit Möglichkeiten der automatischen Erkennung von Informationsstruktur in deutschen Texten.

Begutachtete Publikationen

Grote, Brigitte, Nils Lenke & Manfred Stede (1997) Ma(r)king Concessions in English and German. *Discourse Processes* 24(1), 87-118.

Rösner, Dietmar & Manfred Stede (1992) Customizing RST for the Automatic Production of Technical Manuals. In: R. Dale, E. Hovy, D. Rösner & O. Stock (Hrsg.) *Aspects of Automated Natural Language Generation.* Berlin/Heidelberg/New York: Springer.

Rösner, Dietmar & Manfred Stede (1993) Zur Struktur von Texten – Eine Einführung in die Rhetorical Structure Theory. *KI* 2/93.

Stede, Manfred (1995) Kontrastive Untersuchung einiger kontrastiver Diskurspartikel. *Kognitionswissenschaft* 5(3), 127-140.

Stede, Manfred (1999a) *Lexical Semantics and Knowledge Representation in Multilingual Text Generation.* Dordrecht/Boston: Kluwer.

Kongressbeiträge (zugleich begutachtet)

Grote, Brigitte & Manfred Stede (1998) Discourse Marker Choice in Sentence Planning. In: *Proc. of the Ninth Int'l Workshop on Natural Language Generation*, Niagara/Ontario.

Stede, Manfred (1999b) Rhetorical Structure and Thematic Structure in Text Generation. In: *Proc. of the Workshop on Levels of Representation in Discourse,* Edinburgh University.

Stede, Manfred (2002) Polibox: Generating Descriptions, Comparisons and Recommendations from a Database. In: *Proc. of the 19th Int'l Conference on Computational Linguistics (Coling),* Taipei.

Andere Literatur

Chiarcos, Christian (2002) *Eine Satzplanungskomponente für die Textgenerierung*. Diplomarbeit, FB Informatik, Technische Universität Berlin.

Grabski, Michael (2000) Satztopik und Diskurstopik in Elaboration-Kontexten. In: K. Schwabe, A. Meinunger & D. Gasde (Hrsg.) *Issues on Topics*. *ZAS Papers in Linguistics* 20, 173-207.

Grabski, Michael (2001) Internals from Elaboration. In: O. Teuber & N. Fuhrhop (Hrsg.) *ZAS Papers in Linguistics* 21, 59-66.

3.5 Arbeitsprogramm (Ziele, Methoden, Zeitplan)

3.5.1 Ziele und Methoden

3.5.1.1 Datenbank und Abfragesprachen

Das Projekt soll die Hardware- und Software-Infrastruktur für die zentrale Verwaltung der in den SFB-Teilprojekten erhobenen Korpusdaten bereitstellen und diese dann auch SFB-extern verfügbar machen. Auch die gesprochenen Daten sollen auf dem zentralen Server verwaltet werden, ihre inhaltliche Betreuung sowie die Service-Leistungen bzgl. Annotierungswerkzeugen für gesprochene Daten obliegen allerdings dem Projekt D3. Die grundsätzliche Arbeitsweise soll sein, dass die eigentliche Annotation der Daten dezentral bei den Teilprojekten erfolgt. Technisch sollen alle Daten in XML-Formaten vorliegen, um dann auf den Server gespielt und integriert werden zu können. Bei der Auswahl und Nutzung von Annotationswerkzeugen kann D1 behilflich sein.

Inhaltlich soll im SFB mittelfristig (in enger Abstimmung mit Projekt D2) ein gemeinsamer Annotationsstandard mit IS-Kategorien für die Daten festgelegt werden. Da es aber nicht praktikabel erscheint, sich bereits zu Beginn aller Projekte auf einen solchen Standard zu einigen, sollen die einzelnen Projekte zunächst eigene Kategorien verwenden und der gemeinsame Server eine Reihe von Formaten unterstützen; im Laufe der Zeit werden diese dann in das gemeinsame Schema überführt. Um dies zu realisieren, ist am Anfang eine gründliche Anforderungsanalyse durchzuführen: Bei den Teilprojekten muss festgestellt und dokumentiert werden, welche Art von Daten in welcher Form annotiert werden sollen. Gleichfalls sind existierende Annotationsstandards im Hinblick auf ihre Eignung für die Zwecke des SFB zu evaluieren. Auf der Grundlage dieser Ergebnisse kann dann während der ersten Förderphase zusammen mit D2 ein Annotationsformat entworfen werden, das einerseits den Bedürfnissen der Teilprojekte gerecht wird und andererseits dem „Stand der Kunst“ entspricht. Im engen Kontakt mit den Teilprojekten soll das Annotationsformat ständig evaluiert und weiterentwickelt werden und letztlich in einen „Standard“ münden. Technisch betrachtet wird es sich im Idealfall um eine Erweiterung eines bestehenden (syntaktisch orientierten) Annotationsformats handeln. Die Frage der Integration informationsstruktureller Kategorien in solche Formate stellt eines der ersten Forschungsziele dieses Projekts dar.

Auf dem zentralen Server soll die Software-Umgebung, ähnlich wie im o.g. TIGER Projekt, eine relationale Datenbank (mySQL) gekoppelt mit XML-basierten Austauschformaten sein, die die Schnittstellen zu Nutzern und zu anderer Software bilden. Der Aufbau dieser Umgebung einschließlich der technischen Korpus-Verwaltung wird zu Beginn des Projekts den Arbeitsschwerpunkt bilden.

Existierende Annotierungswerkzeuge werden auf ihre Eignung bzw. hinreichende Erweiterbarkeit überprüft und ggf. wird eine Weiterentwicklung eines solchen Werkzeugs vorgenommen. Für die Datenbank beginnt das eigentliche Einpflegen der Daten (per upload von den Teilprojekten) und die Betreuung. Für die Abfrage der Daten kann im Interesse einer

raschen Verfügbarkeit zunächst eine formale *query language* nach dem Vorbild der in 3.3 genannten Sprachen verwendet werden. Es soll eine WWW-Schnittstelle implementiert werden, die den Zugriff auf die Daten für die Teilprojekte und auch für Nutzer außerhalb des SFB ermöglicht. Auch hier kann auf den Erfahrungen vergleichbarer Projekte aufgebaut werden, um den Entwicklungsaufwand in Grenzen zu halten. Insbesondere ist hier das Werkzeug TASX (Milde & Gut 2002) zu nennen, das nicht nur die Annotation von Sprachdaten auf unterschiedlichen *tiers*, sondern auch deren Bereitstellung per WWW erlaubt. Um mittelfristig die Benutzung der entstehenden Korpora zu erleichtern, soll als Alternative zu den formalen *query languages* eine grafisch orientierte und wissensbasierte Abfragesprache entwickelt werden, die sich den Anforderungen der Nutzer anpassen kann. Grundsätzlich gibt es für Datenbank-Abfrage zwei entgegengesetzte Herangehensweisen: zum einen die genannten formalen Abfragesprachen, zum anderen eine komplett Menü-basierte Umgebung, die das Zusammenstellen eines *query* per Mausklick gestattet. Während der erste Weg das Erlernen einer formalen Sprache voraussetzt, ist der zweite in vielen Fällen zu aufwändig und umständlich. Ein Mittelweg kann in einer intelligenten Unterstützung der NutzerInnen bestehen, indem sie zunächst den allgemeinen Typ der Anfrage angeben und dann bei der Ausgestaltung an jedem Punkt vom System genau die jeweils bestehenden Optionen angeboten bekommen, die auf Wunsch auch erläutert werden können. (Eine Variante dieses Vorgehens wurde von Power & Scott (1998) für das verwandte Problem des Aufbaus von Wissensrepräsentations-Strukturen vorgeschlagen.) Für linguistische Datenbanken wurde ein solcher Zugang unseres Wissens bisher nicht entworfen, erscheint angesichts des vielfältigen Benutzerkreises aber sinnvoll. Das Projekt möchte diesen Weg beschreiten und eine solchermaßen interaktive, System-unterstützte Datenbank-Abfrage ermöglichen.

3.5.1.2 Textstruktur und Textgenerierung

Eine methodische Grundannahme des Projekts ist, dass die tatsächliche Implementierung eines computerlinguistischen Textproduktionsmodells, die einerseits sehr präzise Formulierungen erfordert und andererseits bewertbare Resultate (Texte) liefert, ein sehr nützliches Mittel für die Weiterentwicklung und Validierung theoretischer Beschreibungen darstellt. Das Projekt ILEX (O'Donnell et al. 2001) etwa lieferte dafür ein Beispiel, als anhand des RST-basierten Textplaners das Problem der ungenügenden Berücksichtigung der Wechselwirkungen zwischen rhetorischer Struktur und Fokus-Bewegung zutage trat. Um die Güte eines Produktionsmodells beurteilen zu können – und zudem eine „Werkbank" für den Vergleich unterschiedlicher Lösungen für Teilprobleme zu besitzen – erscheint eine Implementierung daher als sehr vorteilhaft.

Gleichzeitig gilt, dass die vollständige Implementierung eines Textgenerators den Rahmen eines einzelnen Projekts sprengen würde. Daher konzentrieren wir uns auf die für das Projektthema besonders relevante Phase der Satzplanung und lösen andere Aufgaben (d.h. vor allem die grammatische Realisierung) zunächst mit bereits existierenden Komponenten. Von der Aufgabe der inhaltlichen Textplanung soll ganz abstrahiert werden, die Eingabe für das entstehende System sollen deshalb zunächst auf der Basis eines Textkorpus manuell erstellte Textpläne sein. Die System-Architektur soll am Prinzip des in 3.3 skizzierten „Pipeline"-Modells festhalten, jedoch die Aufgabenteilung zwischen Textplanung und Satzplanung neu gestalten, wie unten ausgeführt wird.

Zur weiteren Abgrenzung der Aufgaben soll sich das Projekt zunächst auf eine einzelne Textsorte beschränken, nämlich Kommentare (somit argumentativ orientierte Texte), wie sie in

Tageszeitungen zu finden sind. Ein zweites methodisches Merkmal der Arbeit soll dabei die Betonung korpusbasierten Vorgehens sein. Ausgehend von einem Korpus annotierter Kommentare werden manuell Textpläne erstellt, die dann vom Generierungssystem zu Texten umgesetzt werden, die unmittelbar mit den „Originalen" abgeglichen werden können. Eine Auswertung hinsichtlich der uns interessierenden informationsstrukturellen Aspekte kann Änderungen der Generierungsregeln nahelegen, deren Effekte wiederum unmittelbar kontrolliert werden können, usw.

Im Projektverlauf steht daher die Korpus-Annotation am Anfang. Das Ergebnis, ein hinsichtlich informationsstrukturell relevanter Kategorien annotiertes Korpus, wird nicht nur die Voraussetzung für die nachfolgenden Untersuchungen der Korrelationen zwischen rhetorischer Struktur und IS, sowie für die Arbeiten am Textproduktionsmodell liefern, sondern auch eine eigenständige Ressource bilden, die für die Erforschung anderer IS-Fragen nützlich sein kann.

Korpus-Annotation

Arbeiten zu Diskursstrukturen oder zur Textlinguistik haben in der Vergangenheit oftmals schwer überprüfbare und verallgemeinerbare Resultate geliefert, wofür u.E. die sehr geringe Verfügbarkeit annotierter Text-Daten (und damit einhergehend die Entwicklung von Annotationsstandards) einen wichtigen Grund darstellt. Das Projekt sieht daher die Erstellung solcher Daten als wichtigen Arbeitsschritt. Wie in 3.4 ausgeführt, steht bereits ein Korpus von Zeitungskommentaren zur Verfügung, das hinsichtlich rhetorischer Struktur annotiert wurde; diese Analysen sind zunächst nochmals zu überprüfen und ggf. zu überarbeiten. Parallel zur rhetorischen Struktur sind die Texte dann unter informationsstrukturellen Gesichtspunkten zu annotieren. Technisch werden zwei verschiedene Werkzeuge eingesetzt: Das o.g. RST-Tool, das speziell für die Annotation von Baumstrukturen an Texte entwickelt wurde, sowie ein *tier*-orientiertes Werkzeug (entweder TASX oder EXMARALDA, s.o.) für die IS-Kategorien, die als Labels auf unterschiedlichen *tiers* fungieren. Alle genannten Werkzeuge legen die Daten in XML-Strukturen ab, so dass relativ einfach Skripte erstellt werden können, die systematische Vergleiche zwischen den rhetorischen Bäumen und den IS-*tiers* vornehmen.

Die präzise Ausgestaltung der zu annotierenden Kategorien soll erst zu Beginn der Projektarbeit festgelegt werden; im folgenden wird jedoch der Rahmen beschrieben. Grundlegend ist, dass die Annotationen – da sie erst die Untersuchung und weitere Theoriebildung zur IS befördern sollen – soweit möglich in prä-theoretischer Weise vorgenommen werden, also z.B. noch keinen speziellen Fokus- oder Topik-Begriff voraussetzen. Die Annotationen betreffen zum einen Diskursreferenten und ihre Realisierung, zum anderen auffällige syntaktische Konstruktionen.

Hinsichtlich der Diskursreferenten ist zunächst ihre Ko-Referenz (per Variablenbenennung) zu annotieren, sodann der Grad ihrer „accessibility" im Sinne etwa von Ariel (2001). Sie schlägt eine Skala von 14 Kategorien vor (je höher die linguistische Form auf dieser Skala, desto salienter der Diskursreferent), von obligatorisch elidierten Argumenten über Null-Anapher, verschiedene Arten von Pronomina, hin zu verschiedenen definiten NPs und schließlich der indefiniten NP. Dies ist für die Praxis möglicherweise zu feinkörnig und kann im Sinne der Annotationseffizienz etwas reduziert werden.

Um am Korpus die Voraussetzungen unterschiedlicher „Centering"-Theorien (in der Tradition von Grosz, Joshi, Weinstein (1985)) untersuchen zu können, sollen grammatische Rollen annotiert werden, anhand derer im Zusammenspiel mit der Koreferenz der Centering-Typ des Übergangs (*retain*, *shift*, ...) bestimmt werden kann.

Treten im Text Fokuspartikeln auf, sollen diese mit ihrem Skopus annotiert werden. Die Konnektoren, die rhetorische oder semantische Relationen an der Oberfläche signalisieren (und abhängig von ihrer Position mitunter auch fokus-sensitiv sind, etwa *aber* oder *jedoch*) werden im Zuge der rhetorischen Analyse mit markiert.

Den zweiten Schwerpunkt der Annotation bilden markierte syntaktische Konstruktionen, deren Relevanz in Bezug auf die Informationsstruktur wie auch die rhetorische Struktur diskutiert wurde. So wurde für *it-clefts* von Delin & Oberlander (1995) eine unterschiedliche rhetorische Funktion gezeigt, die von der Deutung der abgespaltenen Konstituente als *topic* oder als *comment* abhängig ist, und die auch für das Deutsche nachvollziehbar ist. Interessant sind ferner „Linksversetzung" (LV) und „Freies Thema", deren versetzten Konstituenten *topic*-Status zugeschrieben wird (Jacobs 2001), und denen u.a. die rhetorische Funktion zugeordnet wird, ein neues Diskurstopik einzuführen (Selting 1993). Ähnliches gilt für „was ... betrifft"-Konstruktionen (engl. *as for constructions*, vgl. (Lambrecht 1994)), die sog. Rahmentopiks spezifizieren. Aus Textproduktionssicht machen diese Konstruktionen auf der Ebene der Satzplanung explizit, welche Referenten in den Aufmerksamkeitsfokus gestellt werden, dort bleiben oder abgelöst werden. Auf der Ebene der Textplanung lassen die beobachteten Textsegmentierungs-Effekte – Beginn, Beibehaltung oder Beendigung eines Diskurstopiks – auf das Vorhandensein von speziellen subordinierenden Diskursrelationen wie etwa *Elaboration* schließen (Asher 1993).

Ein Beispiel für die Relevanz von LV ist ihre Verwendung in einem abschließenden Kommentar zu einem Interview (aus dem NEGRA-Korpus):

> (a) Kein Wunder, daß so einer nichts hören will von einer Krise des Theaters in Deutschland: (b) Das hat nicht nur damit zu tun, daß er selbst in diesem Betrieb mittut. (c) Ja, die Peymanns und Steins und Zadeks, die irgendwann ihre Väter auf dem Theater umgebracht haben, die hätten es wohl versäumt, sich ihre Söhne zu ziehen. (d) Doch Tiefenbacher scheint überzeugt, daß die trotzdem heranwachsen.

(627816: Söhne ohne Väter. Ein Gespräch mit Matthias Tiefenbacher, der im Nachtfoyer „Verlorene Zeit" inszeniert.)

Die rhetorische Struktur dieses Arguments ist grob die, dass zwischen (a), das eine Einstellung des im Interview befragten Regisseurs charakterisiert, und der komplexen Sequenz (b) – (d) die Relation *Explanation* besteht. Die LV in (c) sorgt dafür, dass die in diesem Satz gemachte Aussage der erklärenden Sequenz (als Einschränkung zu der Aussage in (d)) untergeordnet werden kann und ermöglicht damit die Fortsetzung der Sequenz, obwohl bestimmte Inferenzen aus (c) damit nicht vereinbar wären. Die Referenten der linksversetzten Konstituente werden aktiviert, erhalten aber Rollen als „Ausnahmen".

Das entstehende Korpus wird in die zentrale Datenbank integriert, und das verwendete Annotationsformat wird – sowohl technisch als auch inhaltlich (verwendete Kategorien) – mit anderen Teilprojekten, die Texte bearbeiten (vornehmlich B3 und B4) abgestimmt.

Korrelationsanalyse: Rhetorische Struktur und Informationsstruktur

Auf der Grundlage der so annotierten Daten – und fallweise auch mit Blick auf andere Korpora, wenn es um die Erhebung einzelner Phänomene geht – sollen dann (in Zusammenarbeit mit Projekt A3, das ebenfalls rhetorische Struktur, allerdings in gesprochener Sprache, untersucht) Korrelationen zwischen rhetorischer Struktur und Informationsstruktur ermittelt werden. So deutet etwa Lavid (2000) an, dass bestimmte „chaining strategies" auch für argumentative Texte bedeutsam sind, was an unseren Daten überprüft und ggf. spezifiziert

werden soll. Weiterhin seien im Folgenden noch einige Hinweise auf Korrelationen zwischen den beiden Strukturen genannt.

Verwendungen der oben genannten syntaktischen Konstruktionen dienen der Textsegmentierung. Dies kann an Beschränkungen gezeigt werden, die für den Zugriff von Diskursanaphern auf solche Aussagen bestehen, die im Text diesen Konstruktionen vorausgehen. In der rhetorischen Struktur haben solche Beschränkungen eine Entsprechung in der spezifischen Konstruktion der sog. *right frontier* (vgl. Webber 1991).

Daran anknüpfend verfolgt die „Veins Theory" von Cristea et al. (1998) die Hypothese, dass Antezedenten für allgemeine (also nicht nur Diskurs-) Anaphern im vorangehenden Text bevorzugt in den Nuklei der Relationen im RST-artigen Baum gesucht werden. Die rhetorische Struktur, konkret die Verteilung der Nuklei, liefert demnach starke Präferenzen für die Anaphernresolution. Auf eine solche Überlappung zwischen der Wahl referentieller Ausdrücke und der hierarchischen Textstruktur verweist auch Givón (1995, Abschnitt 8.5.4). Eine weitere Ausarbeitung dieser Überlegungen, die im Projekt verfolgt werden soll, könnte ein Distanzmaß zwischen Propositionen vorschlagen, das auf Zahl und Art der intervenierenden Knoten im rhetorischen Baum beruht.

Verwendungen der o.g. syntaktischen Konstruktionen dienen ferner der Realisierung lokaler rhetorischer Struktur unter Umständen, in denen Information auf der Satzplanungsebene angepasst werden muss. Linksversetzung im Korpus-Beispiel oben erlaubt z.B. die Realisierung der Diskursrelation *Explanation*, obwohl die Aussage in (c) Inferenzen zulässt, die nicht zur Erklärung der Aussage in (a) geeignet sind.

Konstruktionen wie die Linksversetzung können lokale rhetorische Struktur ausdrücken (z.B. *Contrast*), mit der auf übergeordnete rhetorische Struktur geschlossen werden kann. Interessant ist in diesem Zusammenhang auch das Ergebnis einer Korpus-Untersuchung von Schilder & Tenbrink (2001), die zeigt, dass englische *before* und *after* clauses, die am Satzbeginn stehen, eine *Elaboration* (im Sinne der SDRT nach Asher (1993)) zum vorangehenden Text induzieren; dies gilt nicht mehr, wenn die clauses an anderer Stelle im Satz stehen.

Zu betrachten sind schließlich auch die Zusammenhänge zwischen rhetorischen Relationen und der Einführung, Fortsetzung und pragmatischen Manipulation von Topikreferenten.

Von der Untersuchung dieser Zusammenhänge sind einerseits Erkenntnisse über Textstrukturierung und eine weitere Explikation der Begriffe *Kohärenz* und *Kohäsion*, andererseits aus Sicht der Automatischen Textgenerierung eine Präzisierung des Begriffes der Satzplanung zu erhoffen.

Textproduktionsmodell

Für die Umsetzung der Erkenntnisse bei der Entwicklung des Textproduktionsmodells ist die Arbeitshypothese, dass die rhetorische Struktur – im Gegensatz zu ihrer Rolle in früheren, RST-inspirierten, Modellen (die die lineare Abfolge der Propositionen oft bereits während der Textplanung komplett festlegten) – eine „tiefe" Repräsentationsebene darstellt, die sich keineswegs unmittelbar an der Textoberfläche niederschlägt. Stattdessen sollen die Satzplanungsprozesse erheblich gestärkt werden, um die Belange der Informationsstruktur berücksichtigen zu können. In der als Textplan fungierenden Repräsentation der rhetorischen Struktur können Propositionen oder Teil-Bäume mit kommunikativen Gewichten (Brandt 1996, s. 3.3) annotiert sein, und sie soll insbesondere die lineare Abfolge von Informations-

einheiten nur partiell festlegen – nämlich insoweit sie unter argumentativen Gesichtspunkten wichtig ist und daher genuin ein Aspekt der Textplanung ist. Präsentiert beispielsweise ein Autor zunächst ein Gegenargument, dann dessen Entkräftung, dann sein eigenes Argument, so handelt es sich um eine rhetorische Figur, deren Abfolge im Textplan festgeschrieben sein muss. Besteht aber z.B. innerhalb des Gegenarguments die Wahl, zunächst den einen oder anderen Aspekt zu behandeln, so kann diese Entscheidung der Satzplanung überlassen bleiben, die dazu Faktoren der thematischen Entwicklung oder „topic continuity" heranziehen kann. Der Textplan soll also hinsichtlich der linearen Abfolge noch unterspezifiziert sein.

Die Satzplanung „konsumiert" Stück für Stück Teile der rhetorischen Struktur und bildet sie auf einen Satzplan, eine lexikalisierte Prozess-Partizipanten-Struktur ab. Bei diesem Schritt der Linearisierung wird zugleich ein „Diskursgedächtnis" aufgebaut, das die relevanten Entscheidungen des bisherigen Textverlaufs (Konstituentenfolge, referierende Ausdrücke, Wortwahlen) zur Verfügung stellt. Diese Information beeinflusst dann ihrerseits die Auswahl des nächsten zu linearisierenden Teils des Textplans, sowie die feineren Entscheidungen innerhalb des entstehenden Satzplans, insbesondere die Konstituentenfolge und die Wahl referierender Ausdrücke. Analog zur Aufgabe der Anaphernresolution bei der Textanalyse sind die bestgeeigneten Formen für das Referieren im Kontext zu finden, wofür das Diskursgedächtnis dann auch die Funktionalität einer DRT-orientierten Repräsentation (Kamp & Reyle 1993) bereitstellen soll. Die Satzplanung macht Vorgaben für die relative Prominenz der einzelnen Konstituenten, die in der letzten Generierungsphase, der Oberflächenrealisierung, durch geeignete syntaktische Mittel umgesetzt werden; an dieser Stelle sind die oben genannten syntaktischen Phänomene, wie die Topikalisierung oder die Linksversetzung, zu behandeln.

Die genaue Ausgestaltung dieses vorgeschlagenen 2-Ebenen-Modells soll der wesentliche Beitrag des Projekts auf der Seite der Textgenerierung sein. Das Modell soll über die derzeit verwendeten starren Abbildungsmechanismen hinausgehen und der Dynamik des Textproduktionsprozesses Rechnung tragen. Durch die Verbindung der „Abarbeitung" des Textplanes mit der Information über die bereits aufgebauten Textstrukturen (Diskursgedächtnis) erhält die Satzplanung eine deutlich stärkere Rolle als in bisherigen Ansätzen und gestattet damit den Übergang zu Textsorten wie Kommentaren, die nur sehr wenig durch „schematischen" Aufbau gekennzeichnet sind (im Gegensatz etwa zu Berichtstexten oder Instruktionen, die bisherige Textgeneratoren mit einfacheren Verfahren produzieren).

Perspektive für die längerfristige Entwicklung: Wenn die beschriebenen textuellen Aspekte behandelt sind, soll sich die Aufmerksamkeit verstärkt auf die Oberflächenrealisierung und damit die Grammatik richten, verschiedene syntaktische Mittel zur Realisierung informationsstruktureller Vorgaben sollen soweit möglich in das Modell integriert werden. In einem weiteren Schritt soll die Arbeit auf gesprochene Sprache ausgedehnt werden, indem von schriftlichen zu gelesenen (Rundfunk-, TV-) Kommentaren übergegangen wird. (Wobei natürlich die Verschiedenheit von gelesener Sprache und Spontansprache zu berücksichtigen ist.) Unabhängig vom Produktionsmodell kann zunächst überprüft werden, ob die entwickelten Vorstellungen von Textstruktur in Zeitungskommentaren auch für die gelesenen Texte gültig sind; hier wird an die Ergebnisse des Projekts A3 anzuknüpfen sein. In Zusammenarbeit mit phonologisch/phonetisch arbeitenden Projekten (D2) könnte dann schließlich eine Ausdehnung des Produktionsmodells auf die Synthese gesprochener Sprache angestrebt werden (was in der Automatischen Textgenerierung unter dem Stichwort „concept-to-speech" derzeit aktuell ist).

3.5.2 Zeitplan

Innerhalb des Gesamtprojekts wird der Schwerpunkt, d.h. auch der Personaleinsatz, in den ersten Monaten auf dem Aufbau der Korpus-Infrastruktur liegen. Wenn diese Funktionalität bereit steht, die Dienstleistungen für die SFB-Teilprojekte angelaufen sind und die Erfordernisse sich hier auf die Sicherstellung des laufenden Betriebs beschränken, sollen die Mitarbeiter sich verstärkt mit den Forschungsfragen zu Korpus-Recherche, Textstruktur und Textgenerierung beschäftigen. Entsprechend wird im Folgenden der Zeitplan für die drei Arbeitsfelder Korpus-Service, Korpus-Forschung und Textproduktionsmodell unter der Maßgabe formuliert, dass die Mitarbeiter zwar primär ihren inhaltlichen Schwerpunkten zugeordnet sind (siehe 3.8.1), aber durchaus auch bei den anderen Aufgaben mitwirken.

	Korpus-Service	**Korpus-Forschung**	**Textproduktionsmodell**
2003	Erfassung der zu erwartenden Daten in den Teilprojekten; Bereitstellung der Hardware- und Software Infrastruktur (Datenbank)	Vergleichende Evaluation existierender Annotationsstandards und Abfragesprachen; erster Vorschlag eines Annotationsstandards für den SFB	RST-Kommentarkorpus überarbeiten, anschließend in Datenbank übernehmen; Festlegung der Kategorien für Informationsstruktur-Annotation; Beginn der Annotationen
2004	Einpflegen von Daten; endgültige Festlegung der Datenformate; Konzeption und Implementierung der WWW-Schnittstelle; Implementierung einer Abfragesprache; Weiterbildung von Teilprojekt-Mitarbeitern; Hard-/Software Betreuung	Festlegung des Annotationsstandards für den SFB; Entwicklung eines Annotationswerkzeugs (durch Ausbau eines existierenden W.); gründliche benutzerorientierte Evaluation existierender Abfragesprachen für linguistische Datenbanken; erster Vorschlag einer Abfragesprache, die Informationsstruktur und Syntax etc. kombiniert	Korpus-Auswertung: Korrelationen zwischen rhetorischer Struktur und Informationsstruktur; Bereitstellung der Software Umgebung für Textgenerierung (Module anpassen); Implementierung des Satzplaners
2005	Evaluation der gewählten Datenformate, ggf. Änderungen; Einpflegen von Daten; Weiterbildung von Teilprojekt-Mitarbeitern; Hard-/Software Betreuung	Evaluation des eigenen Annotationsstandards und des Annotierungswerkzeugs durch Befragung der Teilprojekte, Weiterentwicklung; Evaluation der WWW-Schnittstelle; Implementierung der Abfragesprache	Generierungslexikon für die Lexeme der zu produzierenden Texte; Anpassung des Oberflächenrealisierers; erste Generierungsläufe mit Evaluation
2006	Evaluation der gewählten Datenformate, ggf. Änderungen; Einpflegen von Daten; Weiterbildung von Teilprojekt-Mitarbeitern; Hard-/Software Betreuung	Weiterentwicklung des Annotationsstandards und des Werkzeugs; Evaluation der Abfragesprache, Weiterentwicklung	Weiterentwicklung des Generators; Ausweitung der Text-Abdeckung; Formulierung der Ergebnisse für die Beschreibung von Textstruktur; Test an anderen Textsorten
2007	Evaluation der gewählten Datenformate, ggf. Änderungen; Einpflegen von Daten;	Weiterentwicklung des Annotationsstandards und des Werkzeugs und der Abfragesprache	Experimente mit gelesenen (Radio/TV) Kommentaren; Transkription und Annotation

Weiterbildung von Teilprojekt-Mitarbeitern; Hard-/Software Betreuung	wie im Originalkorpus; Evaluation der Eignung des Annotationsmodells

3.6 Stellung innerhalb des Sonderforschungsbereichs

Das Projekt administriert die zentrale Datenbank, die sich aus den in den Teilprojekten erhobenen Daten speist und die SFB-intern sowie –extern per WWW zugreifbar sein soll. Zudem wird Beratung für die Auswahl und Nutzung von Annotationswerkzeugen angeboten. Eine enge Abstimmung mit D3 wird sicherstellen, dass sowohl geschriebene als auch gesprochene Daten gleichermaßen verwaltet werden. Sowohl hinsichtlich der Daten-Erhebung als auch für die Definition von Annotationsstandards wird D2 ein zentrales Partnerprojekt sein. Mit den Projekten, die Textdaten bearbeiten (B3, B4) ist der Entwurf von IS-Kategorien für die Beschreibungsebene „Text" abzustimmen, mit A3 die Interaktionen zwischen Informationsstruktur und rhetorischer Struktur. Die entstehenden Formate und Standards sollen für den SFB die Vergleichbarkeit und den Austausch der Ergebnisse ermöglichen. Dazu soll das Projekt auch eine einfach verwendbare Abfragesprache für die Daten entwickeln.

Außerhalb des SFB sind die wichtigsten Kooperationspartner: Für Textgenerierung Prof. John Bateman (Uni Bremen), für Fragen der Diskursstruktur Dr. Daniel Marcu (Univ. of Southern California / ISI) und Dr. Maite Taboada (Simon Fraser University).

3.7 Abgrenzung gegenüber anderen geförderten Projekten

Entfällt.

Literatur

Ariel, Mira (2001) Accessibility Theory: An Overview. In: T. Sanders, J. Schilperoord & W. Spooren (Hrsg.) *Text Representation: Linguistic and Psycholinguistic Aspects*. Amsterdam: John Benjamins, 29-88.

Asher, Nicholas (1993) *Reference to Abstract Objects in Discourse*. Kluwer: Dordrecht.

Bateman, John (1997) Enabling Technology for Multilingual Natural Language Generation: The KPML Development Environment. *Natural Language Engineering* 3(1).

Brandt, Margareta (1996) Subordination und Parenthese als Mittel der Informations-strukturierung in Texten. In: W. Motsch (Hrsg.) *Ebenen der Textstruktur*. Tübingen: Niemeyer, 211-240.

Brandt, Margareta & Inger Rosengren (1992) Zur Illokutionsstruktur von Texten. *Zeitschrift für Literaturwissenschaft und Linguistik* 92, 9-51.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius & George Smith (2002) The TIGER Treebank. In: *Proc. of the WS on Treebanks and Linguistic Theories*, Sozopol.

Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski (2001) Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: *Proc. of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Denmark.

Cristea, Dan, Nancy Ide & Laurent Romary (1998) Veins Theory: An Approach to Global Cohesion and Coherence. In: *Proc. of COLING/ACL-98*, Montréal.

Delin, Judy & Jon Oberlander (1995) Syntactic Constraints on Discourse Structure: The Case of It-Clefts. *Linguistics* 33, 465-500.

Fries, Peter (1981) On the Status of Theme in English: Arguments from Discourse. In: S. Petöfi & M. Sozer (Hrsg.) *Micro and Macro Connexity of Texts*. Hamburg: Buske.

Givón, T. (1995) *Functionalism and Grammar*. Amsterdam: John Benjamins.

Grimes, Joseph (1975) *The Thread of Discourse*. The Hague: Mouton.

Grosz, Barbara & Candace L. Sidner (1986) Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12(3), 175-204.

Grosz, Barbara, Aravind Joshi & Scott Weinstein (1995) Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics* 21(2), 203-226.

van Halteren, Hans & Theo van den Heuvel (1990) *Linguistic Exploitation of Syntactic Databases*. Amsterdam: Editions Rodopi.

Hobbs, Jerry (1985) *On the Coherence and Structure of Discourse*. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Jacobs, Joachim (2001) The Dimensions of Topic-Comment. *Linguistics* 39, 641-681.

Kamp, Hans & Uwe Reyle (1993) *From Discourse to Logic*. Dordrecht: Kluwer.

Klein, Wolfgang & Christiane von Stutterheim (1992) Textstruktur und referentielle Bewegung. In: Wolfgang Klein (Hrsg.) *Textlinguistik*. Göttingen: Vandenhoeck & Ruprecht, 67-92

Knott, Alistair, Jon Oberlander, Mick O'Donnell & Chris Mellish (2001) Beyond Elaboration: The Interaction of Relations and Focus in Coherent Text. In: T. Sanders, J. Schilperoord & W. Spooren (Hrsg.) *Text Representation: Linguistic and Psycholinguistic Aspects*. Amsterdam: John Benjamins, 181-196.

Lambrecht, Knud (1994) *Information Structure and Sentence Form*. Cambridge: CUP.

Lavid, Julia (2000) Contextual Constraints on Thematization in Written Discourse. In: P. Bozon, M. Cavalcanti & R. Nossum (Hrsg.) *Formal Aspects of Context*. Dordrecht: Kluwer.

Lezius, Wolfgang (2002) TIGERSearch – Ein Suchwerkzeug für Baumbanken. In: *Proc. der KONVENS-Tagung*, Saarbrücken.

Mann, William & Sandra Thompson (1988) Rhetorical Structure Theory: A Theory of Text Organization. *TEXT* 8(3), 243-281.

Marcus, Mitchell, Beatrice Santorini & Mary Marcinkiewicz (1993) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313-330.

Martin, James (1992) *English Text: System and Structure*. Amsterdam: John Benjamins.

Milde, Jan-Torsten & Ulrike Gut (2002) The TASX Environment: An XML-based Toolset for Time Aligned Speech Corpora. In: *Proc. of the Third Int'l Conference on Language Resources and Evaluation (LREC 2002)*, Gran Canaria.

Moore, Johanna & Cecile Paris (1993) Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics* 19(4), 651-694

O'Donnell, Mick (1997) RST-Tool: An RST Analysis Tool. In: *Proc. of the 6th European Workshop on Natural Language Generation*, Duisburg.

O'Donnell, Mick, Chris Mellish, Jon Oberlander & Alistair Knott (2001) ILEX: An Architecture for a Dynamic Hypertext Generation System. *Natural Language Engineering* 7, 225-250.

Power, Richard & Donia Scott (1998) WYSIWYM: Knowledge Editing with NL Feedback. In: *Proc. of the Ninth Int'l Workshop on Natural Language Generation*, Niagara-on-the-Lake/Ontario.

Ramm, Wiebke & Claudia Villiger (1995) *Global Text Organization and Sentence-Grammatical Realization.* CLAUS Report 61, Universität des Saarlandes, Saarbrücken.

Reiter, Ehud & Robert Dale (2000) *Building Natural Language Generation Systems*. Cambridge: CUP.

Sanders, Ted, Wilbert Spooren & Leo Noordman (1992) Towards a Taxonomy of Coherence Relations. *Discourse Processes* 15.

Schilder, Frank & Thora Tenbrink (2001) Before and After: Sentence-Internal and –External Discourse Relations. In: *Proc. of the WS "From Sentence Processing to Discourse Interpretation"*, Utrecht.

Schiller, A., S. Teufel & C. Stöckert (1999) *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technischer Bericht, Universität Stuttgart und Universität Tübingen.

Selting, Margret (1993) Voranstellungen vor den Satz. *Zs. German. Linguistik* 21, 291-319.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants & Hans Uszkoreit (1997) An Annotation Scheme for Free Word Order Languages. In: *Proc. of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, Washington/DC.

Steiner, Ilona & Laura Kallmeyer (2002) VIQTORYA – A Visual Query Tool for Syntactically Annotated Corpora. In: *Proc. of the Language Resources and Evaluation Conference (LREC)*, Gran Canaria.

Stone, Matthew & Christine Doran (1997) Sentence Planning as Description Using Tree-Adjoining Grammar. In: *Proc. of ACL-97*, 198-205.

Wagner, Andreas & Laura Kallmeyer (2001) Der TUSNELDA-Standard: Ein Korpusannotierungsstandard zur Unterstützung linguistischer Forschung. In: *Proc. der GLDV Jahrestagung*, Gießen.

Wanner, Leo & Eduard Hovy (1996) The HealthDoc Sentence Planner. In: *Proc. of the Eighth Int'l Workshop on Natural Language Generation*, Herstmonceux Castle/GB.

Webber, Bonnie (1991) Structure and Ostension in the Interpretation of Discourse Deixis. *Natural Language and Cognitive Processes* 6(2), 107-135.