Eberhard Karls Universität Tübingen

Sonderforschungsbereich 441: Linguistische Datenstrukturen

Teilprojekt A2: Linguistische Theorien als Datentypen

Antrag für die zweite Phase 2002 – 2004

Prof. Dr. Uwe Mönnich

Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstr. 113
72074 Tübingen
Tel: 07071/2974035
Email: um@sfs.uni-tuebingen.de

3.1 Allgemeine Angaben zum Teilprojekt A2

3.1.1 Thema

Linguistische Theorien als Datentypen

3.1.2 Fachgebiet und Arbeitsrichtung

Formale Linguistik, Texttechnologie

3.1.3 Leiter

Prof. Dr. Uwe Mönnich
geb.Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstr. 113
D-72074 Tübingen
Tel.: 07071-2974035
Fax.: 07071-551335
Email: um@sfs.uni-tuebingen.de

Ist die Stelle des Leiters/der Leiterin des Projektes befristet?

[x] nein [ ] ja, befristet bis zum _____

3.1.4 Aktenzeichen bei bisheriger Förderung in einem anderen Verfahren der DFG

Eine bisherige Förderung in einem anderen Verfahren der DFG liegt nicht vor.

3.1.5 Angaben zu Versuchen und Untersuchungen

In dem Teilprojekt sind vorgesehen

- Untersuchungen am Menschen [ ] ja [x] nein
- klinische Studien im Bereich der somatischen Zell- oder Gentherapie [ ] ja [x] nein
- Tierversuche [ ] ja [x] nein
- gentechnologische Untersuchungen [ ] ja [x] nein

3.1.6 Bisherige und beantragte Förderung des Teilprojektes im Rahmen des Sonderforschungsbereichs

Haushaltsjahr	Personalkosten	Sächliche Verwaltungs-ausgaben	Investitionen	gesamt
1999	136.800,–	2.600,–	—	139.400,–
2000	138.000,–	2.000,–	—	140.000,–
2001	154.800,–	2.000,–	—	156.800,–
Zwischensumme	429.600,–	6.600,–	—	436.200,–
2002	150.600,–	1.000,–	—	151.600,–
2003	150.600,–	1.000,–	—	151.600,–
2004	150.600,–	1.000,–	—	151.600,–

3.2 Zusammenfassung

Daten, insbesondere sprachliche Daten, sind nicht rein gegeben. Die jeweilige Struktur wird durch die theoretische Perspektive, unter der sie betrachtet werden, wesentlich bestimmt.

Das vorliegende Projekt geht in langfristiger Perspektive von der zentralen Hypothese aus, daß die aktuelle Diskussion über die Differenz leitender Paradigmen in der Sprachtheorie die Funktion der zugrunde liegenden Arten von Datentypen unterschlägt und durch die Beschränkung auf den Gegensatz zwischen charakteristischen formalen Verfahrensklassen (generative Erzeugungssysteme, Unifikation, modelltheoretische Constraint satisfaction) den Aspekt der Datenstruktur aus dem Blickfeld verliert. Die mathematischen, wissenschaftssystematischen und praktischen Probleme, die sich aus einer methodisch reflektierten Behandlung des Datenbegriffes ergeben, werden in diesem Projekt behandelt.

In der kommenden, zweiten Phase des Projekts sollen zwei eng miteinander verbundene Schwerpunkte gesetzt werden. Einmal sollen Formalisierungen und Modellierungen der linguistischen Theorien in der chomskyschen Tradition seit 1980 erarbeitet werden. Die mittlerweile vorhandenen drei verschiedenen Teilströmungen, die Rektions- und Bindungstheorie (GB), der Minimalismus und die Optimalitätstheorie, sollen zuerst einzeln untersucht werden. Für die GB soll eine ihrem Charakter als lizensierende Theorie adäquate, bisher nicht existente Modellierung durch Coalgebren erarbeitet werden. Beim Minimalismus, einer generierenden Theorie, soll auf die Arbeiten des ersten Projektleiters in einem früheren SFB zurückgegriffen werden. Schließlich soll die Hypothese analysiert werden, daß die Optimalitätstheorie eine lizensierende Theorie ist. Eine genauere Klärung dieser Fragen würde das Verhältnis von GB zum Minimalismus und zur Optimalitätstheorie aus formaler Sicht beleuchten.

Sie würde insbesondere verdeutlichen, daß die Wahl einer solchen Theorie als Rahmen für linguistische Arbeiten keineswegs arbiträr ist, sondern weitreichende Konsequenzen für die Art und Weise hat, wie der Linguist die Natur einer linguistischen Theorie sieht.

Den zweiten Schwerpunkt bilden berechenbarkeitstheoretische Untersuchungen zu baumähnlichen Strukturen. Die Datenstrukturen, die sowohl linguistischen Theorien als auch semistrukturierten Dokumenten wie annotierten Corpora zugrunde liegen, gehen über Bäume hinaus, ohne gleich die volle Allgemeinheit nichtrestringierter Graphen zu erreichen. Die exakte Analyse der Graphenklassen, in die diese Datenstrukturen fallen, soll einerseits bestehende Unentscheidbarkeitsresultate zur GB-Theorie und zur HPSG komplementieren und andererseits für semistrukturierte Dokumente die Grenze der Komplexität von Annotationen, unterhalb derer Abfragen noch effizient beantwortbar sind, ziehen helfen.

Schließlich ist vorgesehen, die in der ersten Phase begonnene Analyse der Daten der amerindischen Sprache Warao fortzusetzen und die Daten sowie partiellen Analyseergebnisse in ein großes sprachliches Dokumentationssystem zu integrieren und damit der wissenschaftlichen Öffentlichkeit zugänglich zu machen.

3.3 Stand der Forschung

Angesichts des regen Forschungsinteresses, das der allgemeinen Problematik der Datenstrukturen entgegengebracht wird und das sich in einer für den Einzelnen nicht vollständig übersehbaren Fülle an Literatur niederschlägt, wird sich auch dieser Teil der Übersicht zum Stand der Forschung auf ausgewählte Fragestellungen konzentrieren, die mit den im Projekt verfolgten Zielen und den dabei angewandten Methoden in enger Verbindung stehen.

3.3.1 Datenstrukturen und Datenreflexion

Datenstrukturen, insbesondere in ihrer Ausprägung als Mengen, als Algebren und als Theorien, haben in der modernen Sprachwissenschaft in vielfacher Hinsicht eine wichtige Rolle gespielt. Die vertraute Unterscheidung von drei Adäquatheitsebenen aus der früheren Phase des generativen Paradigmas läßt sich unschwer mit den drei Typen von Datenstrukturen in eine methodologische Parallele setzen (Chomsky, 1955/1975). Eine ähnliche Aussage gilt für das Untersuchungsgebiet der formalen Sprachtheorie, in dem es sich als fruchtbar erwiesen hat, die schwache und starke generative Kapazität von Produktionssystemen von der Betrachtung der Komplexität logisch basierter Spezifikationssysteme zu trennen (Thomas, 1997).

Da in zeitgenössischen linguistischen Theorien Typendisziplinen eine wesentliche Funktion erfüllen, muß auch die Entwicklung der Typentheorie in der theoretischen Informatik in einem Forschungskontext berücksichtigt werden, der das Problem sprachli-

cher Datenstrukturen in seinen Mittelpunkt stellt. In der Theorie der Programmiersprachen sind verschiedene Begriffe von Typen herangezogen worden. Typen dienen dazu, Einheiten, die in den jeweiligen Programmierstilen vorkommen, zu klassifizieren. Im Laufe der Entwicklung hat auch hier ein Übergang von den Traditionen, in denen Typen als Mengen von Worten und als sogenannte freie Algebren aufgefaßt wurden, zu einem Konzept stattgefunden, nach dem Typen Modelle von Theorien klassifizieren (Goguen, 1991).

Es ist aufschlußreich, auch den Paradigmenwechsel von generierenden zu lizensierenden Theorien in der Linguistik unter dieser Perspektive der Entwicklung von Datenstrukturen nachzuzeichnen. Unter dem Einfluß der kognitivistischen „Revolution", nach der unsere geistigen Vermögen auf das syntaktische Manipulieren von Symbolen zu reduzieren und mit den Mitteln der Rekursionstheorie zu beschreiben sind, war zunächst die herrschende Auffassung, daß die Frage nach einer adäquaten Charakterisierung der Grammatikalität von Sätzen einer natürlichen Sprache die Frage nach einem Entscheidungsverfahren zwischen wohlgeformten und nicht-wohlgeformten Morphemketten ist (Putnam, 1961). Nachdem es gelungen war, auch auf andere Datenstrukturen neben den natürlichen Zahlen und den Ketten über endlichen Alphabeten induktiv definierte Berechnungsverfahren anzugeben, stellte sich schnell ein innerer Zusammenhang heraus zwischen den scheinbar voneinander unabhängigen Begriffen der Berechenbarkeit, der Induktion und der charakteristischen Eigenschaft der Strukturen, auf die die beiden erstgenannten Begriffe sinnvoll angewandt werden können. Es handelt sich bei dieser Eigenschaft um die entsprechend der jeweiligen terminologischen Tradition als Anarchie, Freiheit oder Initialität bezeichnete Eigenschaft von Strukturen, die insbesondere von der New Yorker Gruppe um Goguen, Thatcher, Wagner und Wright im Rahmen ihrer allgemeinen Theorie abstrakter Datenstrukturen in ihrer Wichtigkeit erkannt wurde (Meseguer und Goguen, 1985).

Eines der Darstellungsmittel, die die weiteste Anwendung zur Repräsentation von linguistischen Strukturen, die im Umkreis von lizensierenden Theorien postuliert werden, gefunden haben, die Merkmalsstrukturen, läßt allerdings diese Eigenschaft der Initialität vermissen (Aczel, 1997). Somit gibt es wichtige technische Gründe, die gegen den Versuch einer Modellierung lizensierender Theorien mithilfe von (initialen) Algebren sprechen. Diese kommen zu den konzeptuellen Gründen hinzu, die in erster Linie in der Einsicht bestehen, daß ein Paradigmenwechsel nur dann angemessen modelltheoretisch abgebildet werden kann, wenn die Datenstrukturen vor und nach dem Paradigmenwechsel mit *prinzipiell unterschiedlichen* Modellierungsverfahren abgebildet werden. Dazu muß ein entsprechendes Modellierungsverfahren für lizensierende Theorien zur Verfügung stehen, und dies war ursprünglich nicht der Fall. Die Situation hat sich jedoch grundlegend geändert, seitdem mit den (finalen) Coalgebren bzw. den mit ihnen äquivalenten „versteckten" Algebren (*hidden algebras*) eine Kategorie gefunden wurde, die die Funktion eines formalen Berechnungsmodells für die Objektorientierung erfüllt (Aczel, 1997; Jacobs, 1996; Reichel, 1995). Bei ihnen handelt es sich um einen versteckten Zustandsraum, über den geeignete Observatoren definiert sind.

Mithilfe der Coalgebren und ihrer Verbindung zu dem objektorientierten Paradigma lassen sich nicht nur die systematischen Beziehungen zwischen verschiedenen Datentypen erfassen, sondern sie bilden außerdem den geeigneten formalen Hintergrund, um Merkmalsstrukturen universal zu charakterisieren (Rutten, 2000). In ihnen sind gewissermaßen beide Entwicklungsstränge, die Differenzierung des Begriffs der Datenstruktur, ausgehend von einer einfachen Klassifikationsmenge in Richtung auf eine Klassifikation von Strukturmodellen durch Theorien, und die Ausbildung eines reflektierten Datenbegriffs als Grundlage der Grammatikforschung, an einem gemeinsamen Ziel zusammengelaufen. Aus diesem Zusammentreffen der skizzierten Entwicklungslinien des Datenbegriffs ergeben sich eine Reihe von Aufgabenstellungen und Hypothesen, unter denen sich einige durch ihre Nähe zur globalen Zielsetzung des Sonderforschungsbereichs und zu den Untersuchungsgebieten bestimmter Einzelprojekte auszeichnen und die Gegenstand des Projekts zum Thema *Linguistische Theorien als Datentypen* sein werden. Im einzelnen handelt es sich dabei um das wissenschaftssystematische Problem des Paradigmenwechsels von generierenden zu lizensierenden linguistischen Theorien und des damit verknüpften Übergangs zu einem anderen Typ von Datenstrukturen und um das formale Problem der Berechenbarkeit von Datenstrukturen.

3.3.2 Berechenbarkeit verschiedener Typen von Datenklassen

Unter dem Aspekt der Berechenbarkeit von Datenstrukturen ist von besonderem Interesse der enge Zusammenhang von algebraischer Struktur und Entscheidbarkeit, der in der sprachwissenschaftlichen Methodologiedebatte nie explizit gemacht worden ist. Dieser Zusammenhang findet auf formaler Ebene seinen Ausdruck in dem Resultat, daß in kategorientheoretischer Sicht Berechenbarkeit und Initialität algebraischer Strukturen zwei Seiten derselben Münze sind (Meseguer und Goguen, 1985; Barr und Wells, 1995). Die algebraischen Voraussetzungen für dieses Resultat laufen in letzter Analyse darauf hinaus, daß sich die relevanten Strukturen in eindeutiger Weise aus Grundelementen und einer Menge von sogenannten Konstruktoren beschreiben lassen. Diese Voraussetzung ist nicht für die Ableitungsbäume erfüllt, die sich kontextsensitiven Produktionssystemen zuordnen lassen, und dieser Umstand liefert unter rekursionstheoretischer Perspektive eine nachträgliche Rechtfertigung für die Zurückweisung kontextsensitiver Grammatiken und die Entscheidung für die Anwendung von Transformationen, die Elemente von eindeutig lesbaren Strukturen in Elemente desselben Typs überführen (Postal, 1964).

Vor diesem Hintergrund des engen Zusammenhangs von algebraischer Struktur und Entscheidbarkeit ist es nicht als zufällig anzusehen, daß im lizensierenden Theorieparadigma das Interesse der Linguisten an Entscheidbarkeitsfragen spürbar abgenommen hat (Chomsky, 1981). Unter den Arbeiten, die sich dennoch mit Entscheidbarkeitsfragen auseinandergesetzt haben, ist für die Prinzipien- und Parameter-Theorie der Ansatz von Rogers (1998) wichtig, weil dieser zeigen konnte, daß die überwiegende Zahl der

postulierten Prinzipien sich im formalen Rahmen der monadischen Logik zweiter Stufe und der regulären Baumfamilien erfassen läßt. Es bleiben allerdings Spezifikationsmechanismen wie z.B. die freie Indizierung übrig, die zur Unentscheidbarkeit führen. Rogers erreicht dieses negative Resultat im Rahmen der starken monadischen Logik, indem er zeigt, wie sich die Indizierung auf das Kachelungsproblem (siehe Börger et al., 1997) reduzieren läßt. Die Frage, ob sich dieses Ergebnis auch schon bei Einschränkung der verwendeten Relationen auf solche, die linguistisch signifikant sind, erzielen läßt, scheint offen zu sein. Bei einer Beantwortung dieser Frage kommt es offenbar in entscheidender Weise darauf an, festzustellen, in welchem Ausmaß durch das Verfahren der Indizierung Baumstrukturen graphenähnlich werden (Seese, 1991, 1992). Das Spektrum der strukturellen Möglichkeiten, innerhalb dessen das Problem angesiedelt ist, läßt sich durch zwei weitere Resultate angeben, die die Komplexität der Berechenbarkeit von bestimmten Problemen kennzeichnen. Wenn keine Strukturen ausgeschlossen werden, dann ist die Relation zwischen der Logik zweiter Ordnung mit einem Existenzquantor als Spezifikationssprache und ihren modellierten Strukturen entscheidbar, genauer: NP-vollständig (Fagin, 1974). Bei einer Beschränkung auf graphenähnliche Strukturen ist dasselbe Verhältnis schon im Hinblick auf die monadische Logik zweiter Stufe unentscheidbar (Courcelle, 1997). Courcelle erzielt sein Resultat wiederum durch eine Reduktion auf das Kachelungsproblem, ein Verfahren, das auch Blackburn und Spaan (1993) anwenden, um die Unentscheidbarkeit einer modallogischen Spezifikationssprache über graphenähnlichen Strukturen zu beweisen. Auf dieses Ergebnis stützen sich wiederum King, Simov und Aldag (1999), um die Unentscheidbarkeit einer bestimmten Formalisierung der HPSG nachzuweisen.

3.3.3 Berechenbarkeit semistrukturierter Daten

Annotierte Dokumente wie etwa ein linguistisches Corpus kann man als semistrukturierte Daten in dem Sinne auffassen, daß es sich bei ihnen weder um einen strukturlosen Wortstrom handelt, noch eine vollständige Strukturierung aller Daten wie etwa bei einer Tabelle einer relationalen Datenbank vorliegt. Eine wichtige Eigenschaft semistrukturierter Daten besteht darin, daß ihre Struktur gebenden Elemente ähnlich wie bei Merkmalsstrukturen über Bezeichner (Namen) angesprochen werden und nicht wie bei Bäumen über ihre Position. Gerade die Tatsache, daß die Daten „nur“ semistrukturiert sind, macht es unmöglich, strukturierende Elemente nur über ihre Position zu bestimmen (Abiteboul et al., 2000). Ein Teilprojekt, das linguistische Datenstrukturen in ihren formalen Eigenschaften untersucht, wird daher natürlicherweise auch auf die Problematik der formalen Analyse semistrukturierter Daten gestoßen, insbesondere da sich einige interessante Parallelen zur Untersuchung der weiter oben beschriebenen linguistischen Datenstrukturen gerade auch bei Berechenbarkeitsfragen ergeben.

Bisherige Untersuchungen zur Komplexität der Verarbeitung von XML-annotierten Dokumenten haben sich in erster Linie auf die Frage konzentriert, wie aufwendig im komplexitätstheoretischen Sinne es ist, die Kompatibilität eines Dokumentes mit ei-

ner vorgegebenen Grammatik (DTD) zu überprüfen (McHugh und Widom, 1999). Es spielen aber auch algebraische Methoden eine Rolle. Eines der Probleme von XML ist die relative Starrheit einer vorgegebenen Grammatik. Diesem Problem, das insbesondere bei der Strukturierung von Antworten auf Anfragen über mehrere Dokumente auftritt, wird mit sogenannten XML-Schemata begegnet, die über mehrere Grammatiken verallgemeinern können. Die Arbeiten von Chidlovskii (2000) und Murata (1999, 2001) zeigen nun, daß XML-Schemata sich sehr gut mit Baumautomaten beschreiben lassen, womit natürlich die entsprechenden Komplexitätsresultate zu Baumautomaten übertragbar werden.

Allen diesen Untersuchungen liegt nach wie vor die Annahme zugrunde, daß jedes annotierte Dokument seiner Struktur nach ein Baum ist. Diese Annahme ist jedoch nur in erster Näherung korrekt. Zwar definiert der Grammatikformalismus für eine DTD, daß jedes Dokument ein Element ist, und daß Elemente nur nebeneinander stehen können oder ein Element ein anderes vollständig beinhaltet, so daß Überlappungen ausgeschlossen sind, und als Grundstruktur ein Baum entsteht. Aber Attribute vom Typ *ID* und *IDREF* erlauben durch Verweise den Aufbau einer weiteren Struktur, die klar über Bäume hinaus zu allgemeinen Graphen geht. Damit werden Komplexitätsresultate der Graphentheorie relevant. Die meisten bekannten Resultate laufen jedoch darauf hinaus, daß das entsprechende Problem entweder direkt unentscheidbar ist oder aber der Lösungsalgorithmus eine solche Komplexität aufweist, daß er bei großen Dokumenten nicht praktikabel ist (van Leeuwen, 1990; Courcelle, 1997). Daher hat sich im letzten Jahrzehnt ein spezieller Zweig der Komplexitätstheorie etabliert, die parametrisierte Komplexitätstheorie, die den nicht handhabbaren Teil schwieriger Probleme isoliert und in einen Parameter „einpackt“, so daß für den verbliebenen Teil wieder Lösungen in der Komplexitätsklasse *PTIME* gefunden werden können. Für dieses Teilprojekt speziell interessant sind dabei Überlegungen in der Graphentheorie zur Frage, wie sich der Übergang von Bäumen zu allgemeinen Graphen gestaltet, in dem Sinne, daß es Klassen von Graphen gibt, die zwar eindeutig über Bäume hinausgehen, aber doch immer noch beschränkt sind, und damit in der parametrisierten Komplexitätstheorie für diese Klassen positive Komplexitätsresultate gewonnen werden konnten (Downey und Fellows, 1999). Dabei spielt insbesondere der Begriff der *Baumweite* eine Rolle, der angibt, wie nahe eine Klasse von Graphen noch den Bäumen ist (Bodlaender, 1993).

3.4 Eigene Vorarbeiten

Der Antragsteller Uwe Mönnich hat sich seit Jahren mit der (co-)algebraischen Modellierung von Grammatikformalismen beschäftigt. In der Arbeit von 1998 kommen kategorientheoretische Prinzipien zur Anwendung, die eine Konzeptualisierung der Baumadjunktionsgrammatiken auf letztlich algebraischer Grundlage gestatten. Die Arbeit von 1999 schlägt vor, die Probleme, die sich jenseits des kontextfreien Bereichs ergeben, mit dem Instrument eindeutiger Homomorphismen aus freien Termal-

gebren zu lösen. Eine algebraische Modellierung des Minimalismus wird in der Arbeit von 2000 vorgenommen. Schließlich hat Mönnich (2001) algebraische Aspekte der lizensierenden Theorie HPSG untersucht. Stephan Kepser hat bereits in seiner Magisterarbeit (1994) den ersten Nachweis geführt, daß die Erfüllbarkeit einer bestimmten Formalisierung der HPSG entscheidbar ist. In der zurückliegenden Phase des Sonderforschungsbereichs (siehe Ergebnisbericht) wurde die HPSG gründlich untersucht. Dabei konnte Kepser (2000) zeigen, daß HPSG als lizensierende Theorie wie gewünscht coalgebraisch modelliert werden kann. Auch wurden komplexitätstheoretische Fragen einer Formalisierung der HPSG in der Arbeit von 2001 behandelt. Frank Morawietz ergänzt mit seinen Arbeiten im Bereich algebraischer Methoden die Resultate zu Coalgebren, die in der ersten Förderphase des Sonderforschungsbereichs erzielt wurden. Neben seinen Veröffentlichungen zusammen mit U. Mönnich stellt insbesondere seine Dissertation (2001) einen Beitrag zu dem Verhältnis von Baumstrukturen verschiedener Komplexitätsklassen dar. Die Erkenntnisse, die Stefanie Herrmann zur Dokumentation einer bedrohten amerindischen Sprache gewonnen hat, stellen eine wichtige empirische Grundlage für die berechnungstheoretische Analyse semistrukturierter linguistischer Daten dar (Herrmann 2001).

3.4.1 Veröffentlichungen

Herrmann, Stefanie (2001): *Demonstrativa im Warao: Bericht einer Feldstudie aus dem westlichen Orinoko -Delta*. Ms, SFB 441, Universität Tübingen.

Kepser, Stephan (1994): *A Satisfiability Algorithm for a Typed Feature Logic*. Sonderforschungsbereich 340, Universität Tübingen, Arbeitsbericht Nr. 60.

Kepser, Stephan (2000): „A Coalgebraic Modelling of Head-Driven Phrase Structure Grammar“, *in* Dirk Heylen, Anton Nijholt, G. Scollo, (Hrsg.), *Proceedings AMiLP2000*, S. 81–95.

Kepser, Stephan (2001): „On the Complexity of RSRL“, *in* Geert-Jan Kruijff, Larry Moss, Richard Oehrle, (Hrsg.), *Proceedings of FG-MOL2001*, ENTCS 53, Kluwer.

Michaelis, Jens, Uwe Mönnich und Frank Morawietz (2000): „Algebraic Description of Derivational Minimalism“, *in* Dirk Heylen, Anton Nijholt, G. Scollo, (Hrsg.), *Proceedings AMiLP2000*, S. 125–141. In Verbindung mit SFB 340 A8.

Michaelis, Jens, Uwe Mönnich und Frank Morawietz (2001): „On Minimalist Attribute Grammars and Macro Tree Transducers“. *Erscheint in* Christian Rohrer, Antje Rossdeutscher und Hans Kamp, (Hrsg.), *Linguistic Form and its Computation*, CSLI; im Druck.

Mönnich, Uwe (1998): „TAGs M-Constructed“, *in Proceedings of the TAG+ Workshop*, Philadelphia.

Mönnich, Uwe (1999): „On Cloning Context-Freeness“, *in* Hans-Peter Kolb und Uwe Mönnich, (Hrsg.), *The Mathematics of Syntactic Structure*, Mouton de Gruyter.

Kolb, Hans-Peter und Uwe Mönnich, (Hrsg.) (1999): *The Mathematics of Syntactic Structure: Trees and Their Logics*, Mouton de Gruyter.

Kolb, Hans-Peter, Jens Michaelis, Uwe Mönnich und Frank Morawietz (2000): „An Operational and Denotational Approach to Non-Context-Freeness". Erscheint in *Theoretical Computer Science*, Elsevier.

Kolb, Hans-Peter, Uwe Mönnich und Frank Morawietz (2000): „Descriptions of Cross-Serial Dependencies". Erscheint in *Grammars*.

Mönnich, Uwe (2001a): „Model-Theoretic Description of TAGs", *in* Geert-Jan Kruijff, Larry Moss, Richard Oehrle, (Hrsg.), *Proceedings of FG-MOL2001*, ENTCS 53, Kluwer.

Mönnich, Uwe (2001b): *Regular Description of HPSG*. Ms, SfS, Universität Tübingen.

Morawietz, Frank (2001): *Two-Step Approaches to Natural Language Formalisms*. Dissertation, SfS, Universität Tübingen.

3.5 Ziele, Methoden, Arbeitsprogramm und Zeitplan

3.5.1 Ziele und Methoden

Den Mittelpunkt des Projekts bildet weiterhin die Hypothese, daß aus einer systematischen Analyse der vorausgesetzten Datentypen Aufschlüsse über Paradigmenwechsel in der Sprachtheorie zu erzielen sind. Es werden zwei Klassen von Datenstrukturen in der Absicht unterschieden, sie als charakteristisch für generierende und lizensierende Ansätze linguistischer Theoriebildung nachzuweisen. Bei den zwei Arten von Datentypen, deren Fruchtbarkeit für eine grundlegende Differenzierung linguistischer Erkenntnisobjekte postuliert wird, handelt es sich um (freie) Algebren und (finale) Coalgebren. Soweit die formalen Eigenschaften dieser Strukturen Gegenstand der Untersuchung sind, werden die Methoden aus den einschlägigen Gebieten der Logik, der Universellen Algebra, der Theorie der Coalgebren und der Komplexitätstheorie angewandt. Auf rein formalem Weg sollen Behauptungen überprüft werden, die auf der Basis kognitivistischer Annahmen über die Berechenbarkeit verschiedener Modelle linguistischer Beschreibungen aufgestellt worden sind, wobei im Sinne der Leitidee des Projekts eine solche Überprüfung ihren Ausgang von den modellspezifischen Datentypen zu nehmen hat. Ein Versuch, der sich die Bestätigung einer Annahme über die Affinität von Theorie- und (abstrakten) Datenklassen zum Ziel gesetzt hat, kann sich nicht auf das Arsenal formaler Instrumentarien beschränken, sondern wird durch eine wissenschaftssystematische Analyse der Konsequenzen, die sich aus den Axiomen der jeweiligen linguistischen Forschungsprogramme ergeben, zu ergänzen und zu relativieren sein. Eine Deckungsgleichheit der Identitätskriterien für Daten- und Theorieklassen ist nicht zu erwarten, und eine wichtige Aufgabe wird darin bestehen, die

Sprengkraft entdeckter Divergenzen zwischen Formen des linguistischen Erkenntnisgegenstands und Formen des theoretischen Zugriffs zu bewerten.

Insbesondere für lizensierende Theorien soll untersucht werden, inwieweit der durch die Art der Theoriebildung vorgegebene Übergang zu Graphen tatsächlich vollzogen ist und vollzogen werden muß. Mit Methoden der Graphentheorie werden wir dabei der Frage nachgehen, ob sich geeignete spezielle Klassen von Graphen finden lassen, die besser als allgemeine Graphen den Intuitionen der Linguisten entsprechen und bessere komplexitätstheoretische Eigenschaften besitzen.

Es sei an dieser Stelle ausdrücklich darauf hingewiesen, daß die Beschäftigung mit graphenähnlichen Strukturen keine wesentliche Einschränkung im Hinblick auf die methodischen Intentionen lizensierender Theorien darstellt. Ein kennzeichnendes Merkmal dieses Theorieparadigmas besteht in der Definition des Grammatikalitätsbegriffs im Rahmen der klassischen modelltheoretischen Erfüllbarkeitsrelation, wie sie von Tarski für die Prädikatenlogik abschließend bestimmt worden ist. Relationale Strukturen, in denen die Sprache der Prädikatenlogik interpretiert wird, stellen auf den ersten Blick eine Verallgemeinerung gegenüber graphenähnlichen Strukturen dar. Die Einsicht, daß sich beliebige relationale Strukturen als Graphen in effizienter Weise kodieren lassen, gehört allerdings zu dem Arsenal „folkloristischer“ Techniken der endlichen Modelltheorie (Ebbinghaus und Flum, 1995). Modulo einer entsprechenden Kodierung lassen sich daher relationale Strukturen unter berechnungstheoretischer Perspektive als Graphen betrachten.

Die Analyse der zugrunde liegenden Datentypen spielt auch für Annotationssprachen wie XML eine wichtige Rolle. Das Projekt verfolgt die Hypothese, daß die in semistrukturierten Dokumenten vorfindbaren Datenstrukturen zum Teil ähnliche Eigenschaften besitzen wie jene, die manchen lizensierenden Theorien zugrunde liegen, insofern als daß semistrukturierte Dokumente zwar primär als Bäume aufgefaßt werden, die Ausdrucksmöglichkeiten einer Annotationssprache wie XML aber so stark sind, daß die Klasse der echten Bäume verlassen wird. Die Frage, ob sich zur Modellierung der Datentypen geeignete spezielle Unterklassen der Graphen finden lassen, spielt für Annotationssprachen eine ungleich größere Rolle, da die mit den Definitionen der Graphenklassen verbundenen Komplexitätsresultate unmittelbare Auswirkung auf die Möglichkeit, strukturierte Anfragen an semistrukturierte Dokumente zu stellen, haben.

3.5.2 Arbeitsprogramm

Zur theoretischen Relevanz von Datenstrukturen

Die zentrale Hypothese des Projekts zieht eine Parallele zwischen den verschiedenen Paradigmen der generierenden und lizensierenden Theoriebildung und den formalen Eigenschaften der sie kennzeichnenden Datentypen. Für das Paradigma der lizensierenden Theorievariante wird eine Form der Gegenstandskonstitution postuliert, die mit den Techniken der coalgebraischen Systemtheorie präzisiert zu werden vermag.

Diese theoretischen Vorgaben im Zusammenhang mit der der Systemtheorie zugewiesenen Rolle haben jene Teile des Arbeitsprogramms zur Folge, die den theoretischen Aspekt von (Typen von) Datenstrukturen berühren. Insbesondere handelt es sich dabei um den Nachweis, daß ein zentrales intendiertes Modell lizensierender Theorien, die Merkmalsstrukturen, mit den Mitteln der Systemtheorie beschreibbar sind, und um die Diskussion der „richtigen“ Auffassung von Merkmalsstrukturen.

Da die Begriffe, mit denen die Systemtheorie arbeitet, relativ neu sind, sei zumindest der Versuch gemacht, die Behandlung der Merkmalsstrukturen in der Hoffnung, daß die folgenden Bemerkungen dennoch eine Idee von der geplanten Durchführung des Arbeitsprogramms vermitteln, zu erläutern. Es sei daran erinnert, daß sich binäre, etikettierte Bäume durch die Angabe von zwei „Konstruktoren“ beschreiben lassen: Ein solcher Baum ist entweder leer oder trägt ein Etikett an seiner Wurzel und verzweigt sich in zwei Unterbäume. Diese Form der algebraischen Spezifikation hat ein kanonisches Modell, das aus den geschlossenen Termausdrücken konstruiert werden kann, d.h. den Ausdrücken, die sich durch iterative Anwendung der Konstruktoroperationen rekursiv generieren lassen. Diese Menge der Termausdrücke ist isomorph zu den endlichen Bäumen. Dual zu diesem Vorgehen beruht die coalgebraische Spezifikation auf sogenannten Destruktoroperationen. Sie beinhalten, welche Operationen oder Beobachtungen auf dem Datentyp der Bäume vorliegen, nämlich die Entfernung eines Etiketts oder die Wahl einer rechten oder linken Verzweigung an einem Knoten. Diese Beobachtungen machen keine Aussagen über die interne Struktur der „Baumwelt“ und erlauben zu ihr nur einen beschränkten Zugang über die mit den Destruktoren verbundenen Manifestationen. Der entscheidende Unterschied zwischen der algebraischen und der coalgebraischen Spezifikation liegt darin beschlossen, daß in dem einen Fall die Operationen in die Baumwelt hineingehen, während sie in dem anderen Fall die umgekehrte Richtung haben. Diesem Umstand entspricht, daß das kanonische Modell der algebraischen Spezifikation eine initiale Algebra ist, während das kanonische Modell der coalgebraischen Spezifikation eine finale Coalgebra bildet.

In der ersten Phase des Projekts konnte gezeigt werden, daß sich die Head-Driven Phrase Structure Grammar (Pollard und Sag, 1987, 1994) mit coalgebraischen Methoden erfolgreich modellieren läßt (siehe Ergebnisbericht). In der nun folgenden Phase möchten wir uns dem Feld der Generativen Grammatik in der chomskyschen Tradition zuwenden und die Entwicklungen seit Beginn der 1980er Jahre betrachten. Seit der Ablösung der Transformationsgrammatiken der 1960er und 1970er Jahre gibt es in der chomskyschen Tradition mindestens die drei Theorieströme Rektions- und Bindungstheorie (GB-Theorie), Minimalismus und Optimalitätstheorie. Trotz einiger Bemühungen zur Formalisierung der GB-Theorie (Chomsky, 1981, 1986; von Stechow und Sternefeld, 1988) gibt es unserer Kenntnis nach keine formale Modellierung, die dem Charakter der GB-Theorie als lizensierender Theorie gerecht wird. Auch die Arbeit von Rogers (1998) behandelt durch die Verwendung algebraischer Methoden die GB-Theorie eher als generierende Theorie. Dies ist ein umso größeres Defizit, als den Autoren der GB-Theorie, allen voran Chomsky, durchaus bewußt war, daß

sich hier im Vergleich zu den bisherigen Transformationsgrammatiken ein Paradigmenwechsel vollzog, was sich an entsprechenden Zitaten von Chomsky (1981, S. 3) gut belegen läßt. Zu einem grundlegenden Verständnis der formalen Eigenschaften der GB-Theorie ist daher eine konzeptionell adäquate Formalisierung auch heute noch erforderlich, obwohl die gegenwärtige linguistische Diskussion sie selbst kaum noch als Rahmen verwendet. Darüber hinaus bildet die GB-Theorie natürlich die Grundlage sowohl für den Minimalismus als auch für die Optimalitätstheorie, so daß für ein Verständnis der formalen Eigenschaften von Minimalismus und Optimalitätstheorie eine Modellierung der GB-Theorie als Basis und Vergleichspunkt erforderlich ist. Da die GB-Theorie eine lizensierende Theorie ist, bietet sich zumindest der Versuch einer coalgebraischen Modellierung an, wobei aber vermutlich erst überprüft werden müßte, ob die allgemeine Theorie der Coalgebren in geeigneter Weise um Relationen erweitert worden ist oder werden kann. In diesem Rahmen ist auch zu klären, inwieweit es sich bei den zugrunde liegenden Datenstrukturen tatsächlich um die von den Linguisten postulierten Bäume handelt, oder ob diese nicht soweit erweitert wurden, daß eher Merkmalsstrukturen vorliegen. Schließlich sind die Knoten in einem Baum längst komplexe Merkmalsbündel. Auch führen eine lizensierende Interpretation des X-Bar-Schemas und die freie Indizierung aus der Klasse der Bäume hinaus.

Der Minimalismus (Chomsky, 1995) stellt in mancher Hinsicht eine Rückkehr zu transformationellen Sichtweisen dar und ist als generierende Theorie zu sehen. U. Mönnich hat sich im Rahmen eines früheren Projekts (SFB 340 A8) bereits intensiv mit einer algebraischen Analyse des Minimalismus auseinandergesetzt (Michaelis, Mönnich und Morawietz, 2000). Diese soll im vorliegenden Projekt für einen Vergleich mit der GB-Theorie und der Optimalitätstheorie herangezogen werden.

Die Optimalitätstheorie (Prince und Smolensky, 1993; Barbosa et al., 1998) ist zwar von ihrer Ausrichtung her eine lizensierende Theorie wie die GB-Theorie, da aber aus den Kandidatenstrukturen die optimale in *mehreren* Schritten ausgefiltert wird, ist nicht zu erwarten, daß der zur Modellierung der GB-Theorie verwendete Ansatz hier einfach übernommen werden kann. Ob eine mehrstufige Ausfilterung mit coalgebraischen Methoden angemessen modellierbar ist, soll im Projekt ergründet werden.

Zur Berechenbarkeit verschiedener Typen von Datenklassen

Mit der ursprünglichen Definition von Komplexitätsklassen auf der Grundlage der Chomsky-Hierarchie war im generierenden Theorieparadigma die Erwartung verbunden, daß mit diesen Klassen wesentliche Eigenschaften der Struktur natürlicher Sprachen erfaßt seien. Diese Erwartung war von der Idee inspiriert, daß die natürlichen Sprachen nur einen generativen Mechanismus begrenzter Ausdrucksstärke benötigten und sie auf diesem Weg von einer reichen Klasse nicht-natürlicher Sprachen in empirisch signifikanter Weise abgegrenzt werden könnten. Dieses Forschungsprogramm wurde in der Theoriebildung aufgegeben, die Sprachen über die Angaben von Beschränkungen definiert, die von Strukturanalysen der Ausdrücke dieser Sprachen zu

erfüllen sind. Der Grund für die Aufgabe des ursprünglichen Forschungsprogramms ist auf der einen Seite in der Tatsache zu sehen, daß die zur Bestimmung der Komplexität lizensierter Modellklassen notwendigen Methoden teilweise erst entwickelt werden mußten und daß auf der anderen Seite mit dem Paradigmenwechsel in der Theoriebildung ein prononciertes Desinteresse an Fragen generativer Kapazität von Formalismen einherging, das von der Überzeugung gespeist war, mit der Reduktion des Spektrums natürlicher Sprachen auf endlich wenige Klassen deren rekursionstheoretische Eigenschaften vernachlässigen zu können (Chomsky, 1981).

Vor diesem forschungsstrategischen Hintergrund ist der Nachweis von Rogers (1998), daß die Rektions- und Bindungstheorie wahrscheinlich eine entscheidbare Formalisierung gestattet, von besonderem Interesse und war der Auslöser für eine Reihe forschungsintensiver Unternehmen mit dem Ziel, die von Rogers gelassenen Lücken zu schließen. Die von ihm als Grundlage gewählte Theorie endlicher Bäume entspricht mit den von ihr lizensierten Modellklassen genau den kontextfreien Sprachen, was zu unmittelbaren Problemen mit der Behandlung der empirisch gut abgesicherten Kreuzabhängigkeiten in Sprachen wie Bambara, Holländisch oder Schweizerdeutsch führt. Die andere Lücke hängt mit der freien Indizierung zusammen, die, zumindest in der Darstellung von Rogers, die Entscheidbarkeit der von ihm zugrunde gelegten formalen Theorie sprengt.

Ein Schwerpunkt in diesem Bereich des Arbeitsprogramms wird daher in dem Versuch bestehen, die von Rogers hinterlassenen Probleme, die mit der komplexitätstheoretischen Einschätzung lizensierender Theorien verbunden sind, einer formal abgesicherten und linguistisch korrekten Lösung zuzuführen. Im Fall der Erweiterung des kontextfreien Bereichs sei ein möglicher Ausweg skizziert. Es ist seit Montagues Vorschlägen zu einer algebraisch formulierten Syntax bekannt, daß die generative Kapazität neben dem Regelformat in entscheidendem Maße von den syntaktischen Operationen abhängt, die von einer Grammatiktheorie zugelassen werden, eine Einsicht, die in virtuoser Weise im kategorialgrammatischen Umfeld umgesetzt wird. Formuliert man die Regeln in der Weise, daß diese Operationen zunächst nur benannt und ihre Ausführung einem weiteren Schritt überlassen wird, entstehen als Ergebnis der ersten Phase dieses schrittweisen Verfahrens zunächst nur reguläre Bäume. Die Vermutung ist nun, daß die Ausführung jener Klasse von Operationen, die für die Darstellung der Kreuzabhängigkeiten ausreicht, wieder auf der Basis jener Theorie möglich ist, von der Rogers ausgegangen ist. Sollte sich diese Vermutung als richtig erweisen, würde die Form der bisherigen Theoriebildung nicht verlassen, sondern nur zu einem zweistufigen Ansatz verfeinert: In der ersten Stufe werden mit den Mitteln derselben Theorie endlicher Bäume die regulären Baumfamilien definiert, und auf der zweiten Stufe werden unter Ausnutzung desselben definitorischen Verfahrens diese Baumfamilien in eine Konfiguration überführt, die die Kreuzabhängigkeiten in deskriptiv angemessener Weise repräsentiert. Es existieren bereits weitreichende Vorarbeiten zu diesem Problem von Mönnich und Morawietz (Kolb, Michaelis, Mönnich und Morawietz, 2000; Michaelis, Mönnich und Morawietz, 2001), in denen dieser zweistufige Ansatz

jeweils für bestimmte Grammatiken realisiert wird. Diese sind jedoch Beschreibungen für spezielle Grammatiken und keine Charakterisierungen, für welche Typen von Grammatiken ein solcher Ansatz verwendbar ist. Neben der offenen Frage der Komplexität dieses Ansatzes stellt sich darüber hinaus auch das konzeptuelle Problem, ob die verwendeten Vereinfachungen, die aus der Fragestellung im wesentlichen ein Kodierungsproblem machen, in dieser Form auch zulässig sind.

In beiden lizensierenden Theoriewelten, sowohl in der GB-Theorie als auch in der Head-Driven Phrase Structure Grammar, kann man einen Unterschied zwischen der allgemeinen Definition der Datenstruktur des zugrunde liegenden Formalismus und den in konkreten Fällen von Linguisten verwendeten Strukturen dahingehend beobachten, daß die allgemeine Definition generellere Strukturen zuläßt, als tatsächlich zur Anwendung kommen. Die Unentscheidbarkeitsresultate von Rogers (1998) für die GB-Theorie und von King et al. (1999) für die HPSG beziehen sich jeweils auf die allgemeine Definition. Ein Teil des Arbeitsprogramms wird daher darin bestehen, diese Unentscheidbarkeitsresultate dahingehend zu komplementieren, daß die jeweils schwächeren tatsächlich verwendeten Datenstrukturen komplexitätstheoretisch untersucht werden, in der Hoffnung, hier zu Entscheidbarkeitsresultaten zu kommen. Im Falle der HPSG sind die Chancen dazu jedoch aufgrund des starken Unentscheibarkeitsresultats von Kepser (2001) (siehe Ergebnisbericht) nicht sonderlich hoch einzuschätzen, während bei der GB-Theorie, wie oben skizziert, durch den Versuch der Reformulierung der Indizierungsmöglichkeiten ein Weg vorgezeichnet ist. In beiden Fällen wird untersucht, wie baumähnlich die verwendeten Datenstrukturen tatsächlich sind. Dazu wird der aus der Graphentheorie stammende Begriff der *Baumweite* (siehe Bodlaender, 1993) benutzt, die relative graphentheoretische Nähe zu Bäumen zu bestimmen. Auf der Basis dieses Wertes sollen dann Standardergebnisse der Graphentheorie die Entscheidbarkeitsfrage beantworten. Im Falle eines positiven Resultates ist sogar noch mehr zu erwarten. Die besonders in der Graphentheorie wichtige parametrisierte Komplexitätstheorie (siehe Downey und Fellows, 1999) ermöglicht unter Umständen Aussagen zur Komplexität der Entscheidbarkeit. Voraussetzung für eine erfolgreiche Übertragung dieser allgemeinen komplexitätstheoretischen Resultate ist die Identifikation linguistischer Datenstrukturen als ganz bestimmte, möglichst baumähnliche, Klassen von Graphen, für die entsprechende Resultate vorliegen.

Zur Berechenbarkeit semistrukturierter Daten

Bemerkenswerterweise stellt sich die Frage nach der Baumähnlichkeit der Datenstrukturen auch bei durch Annotation (teil-)strukturierten Dokumenten. Zwar sind Annotationssprachen wie SGML und XML so angelegt, daß ihre Elemente jeweils paarweise als Anfangs- und Endmarken vorkommen und sich Elemente nicht teilweise überlappen dürfen, also ein Element vollständig Teilelement eines anderen ist oder überschneidungsfrei daneben steht, so daß die Basisdatenstruktur Bäume sind. Aber Verweismöglichkeiten, wie sie etwa durch die XML-Attributtypen *ID* und *IDREF*

gegeben sind, bieten eine Möglichkeit zum Aufbau von über reine Bäume hinausgehenden Datenstrukturen. Damit stellt sich erst einmal die Frage, welche Klassen von Graphen mit Annotationssprachen wie XML ausdrückbar sind. Sollte sich die Anfangshypothese, daß die gesamte Klasse der gewurzelten gerichteten Graphen ausdrückbar ist, als richtig erweisen, hätte dies wichtige komplexitätstheoretische Konsequenzen für die Abfrage an eine Sammlung annotierter Dokumente. Bei Fragen der Komplexität einer Abfrage an eine Sammlung annotierter Dokumente steht meist die Ausdrucksstärke der Abfragesprache als Parameter im Vordergrund. Tatsächlich spielt aber auch die zugrunde liegende Datenstruktur eine wichtige Rolle. Je nach Datenstruktur der Dokumente kann die Komplexität erheblich variieren. Wie oben bereits gesagt, ist etwa die Monadische Theorie zweiter Stufe über der Klasse der Bäume entscheidbar, während sie über der Klasse der (allgemeinen) Graphen unentscheidbar ist. Findet die obige Anfangshypothese Bestätigung, so folgt daraus, daß selbst für relativ einfache Abfragesprachen Unentscheidbarkeitsresultate erzielt werden. Dabei ist andererseits offensichtlich, daß bei der Abfrage großer Corpora ein lediglich positives Entscheidbarkeitsresultat überhaupt nicht ausreicht. Es ist vielmehr erforderlich, daß die Anfrage in polynomialer Zeit abgearbeitet werden kann, da sonst zumutbare Antwortzeiten nicht erreicht werden können. Ein Teil des Arbeitsprogramms wird demnach darin bestehen, möglichst baumähnliche Klassen von Graphen zu identifizieren, für die bei für Benutzer interessanten Abfragesprachen polynomiale Komplexität erreicht wird, sowie entsprechende Einschränkungen für Annotationssprachen anzugeben, um zu gewährleisten, daß die den auf diese Weise annotierten Dokumenten zugrunde liegenden Datenstrukturen in den entsprechenden Graphenklassen liegen. Dazu ist natürlich auch zu klären, welche Abfragesprachen von Benutzern benötigt werden.

Das Problem effizienter Abfragesprachen ist von unmittelbarer praktischer Relevanz, da mit der Gründung der Open Language Archives Community (OLAC, *http://www.language-archives.org*) im Dezember 2000 der Grundstein für ein linguistisches Resourcenarchiv erheblichen Umfangs gelegt ist. Nach den bisherigen Vereinbarungen zeichnet sich ab, daß die Daten gemäß dem oben skizzierten semistrukturierten Format abgelegt werden. Zwar ist es möglich, die entsprechend annotierten Daten in Mengen von relationalen Tabellen zu transformieren, und auf diese Tabellen dann die Techniken relationaler Abfragesprachen anzuwenden, ein Verfahren, das von dem Projekt A1 favorisiert wird. Der Vorteil eines Rückgriffs auf die bewährte Praxis des relationalen Konzepts besteht in der hohen Effizienz von Sprachen, die nach dem Muster von SQL aufgebaut sind. Der Nachteil aber besteht in der systembedingten eingeschränkten Ausdrucksstärke. In vielen Fällen handelt es sich gerade in linguistisch interessanten Zusammenhängen um Strukturen, die nur partiell spezifiziert sind, und die darüberhinaus rekursiv eingebettet und eventuell sogar zyklisch angelegt sein können, Eigenschaften, die von dem Kalkül und den Algebren, die in dem klassischen Datenbankkontext entwickelt wurden, nicht erfaßt werden. Um für diese Eigenschaften ein formal adäquates Ausdrucksmedium zu schaffen, wird in jüngster

Zeit auf das Konstrukt von erweiterten Pfadausdrücken zurückgegriffen, das die Ausdrucksstärke von Abfragesprachen wesentlich erweitert (Abiteboul, 1997; Neven und Schwentick, 2000). Die Untersuchung von Pfadausdrücken spielt in der Entwicklung des semistrukturierten Datenmodells eine wichtige Rolle, und im Kontext der Projektarbeit zur Berechenbarkeit dieses Datenmodells soll diese Fragestellung mit den Mitteln der monadischen Logik aufgegriffen werden.

Zur deskriptiven Relevanz von Datenstrukturen

Wie bereits im Ergebnisbericht dargestellt, hat sich eines der ursprünglichen Projektziele, nämlich in einer prototypischen Arbeit den Wert einer Modellierung sprachlicher Daten samt ihres kulturellen, soziologischen und Umweltkontexts mit objektorientierten Systemen nicht realisieren lassen, da die dazu in Aussicht genommenen Dokumentationssysteme LinguaLinks und EUDICO entweder nicht weiterentwickelt wurden (LinguaLinks) oder bisher nur in einem unvollständigen, normalen Benutzern kaum zumutbaren Zustand vorhanden sind (EUDICO). Da eine wesentliche Verbesserung dieser Situation in der kommenden Phase nicht zu erwarten ist, sehen wir keine Möglichkeit, die mit diesem Projektteil ursprünglich verbundenen theoretischen Intentionen weiter zu verfolgen. Andererseits soll aber auch die in der zurückliegenden Phase von St. Herrmann geleistete Arbeit in der ethnolinguistischen Aufarbeitung der von ihr gesammelten Daten zur amerindischen Sprache Warao nicht vollständig aufgegeben werden. Daher wird ein Teil des Arbeitsprogramms darin bestehen, diese Aufarbeitung fortzusetzen. Dabei werden die Daten einerseits so ausgewählt werden, daß aus einem einfachen Wörterbuch, Tonaufnahmen, transkribierten Texten und Corpora von Mythensammlungen eine für andere Linguisten brauchbare Datenbank des Warao entsteht. Selbstverständlich wird durch die Verwendung allgemein verbreiteter Datenbanken und Annotationsformate sichergestellt werden, daß eine spätere Integration in EUDICO oder besser DOBES, dem Großprojekt zur Dokumentation bedrohter Sprachen (*http://www.mpi.nl/world/DOBES*), problemlos möglich sein sollte. Andererseits sollen die gewonnen Daten auch auf spezielle Fragestellungen, die in den phänomenorientierten Projekten des SFB bearbeitet werden, untersucht werden. So wird unter anderem das System der Deiktika weiter analysiert werden.

3.5.3 Zeitplan

1. Jahr Arbeit an folgenden Schwerpunkten:

- Beginn der Untersuchungen zu Entscheidbarkeitsfragen graphenähnlicher Strukturen.
- Beginn der coalgebraischen Modellierung der Rektions- und Bindungstheorie.
- Analyse und Dokumentation ausgewählter Phänomene des Warao.

2. Jahr

- Fortsetzung der Untersuchungen zu Entscheidbarkeitsfragen mit besonderem Augenmerk auf semistrukturierte Daten.
- Fortführung der coalgebraischen Modellierung der Rektions- und Bindungstheorie.
- Fortsetzung der Analyse des Warao und Vorbereitung der Integration in Dokumentationssysteme.

3. Jahr

- Zusammenfassung der Ergebnisse zu formalen Eigenschaften von Datentypen.
- Modellierung der Optimalitätstheorie und Vergleich mit der Modellierung der GB-Theorie und des Minimalismus.
- Abschluß und evtl. Integration der deskriptiven Ergebnisse zum Warao in große Dokumentationssysteme.

3.6 Stellung innerhalb des Sonderforschungsbereichs

Linguistische Datenstrukturen lassen sich unter verschiedenen Gesichtspunkten zu Einheiten zusammenfassen. Daten, die zu einem Phänomenbereich gehören, die einem Areal zuzuordnen sind, die einer historischen Phase entstammen oder den Sprachgebrauch einer sozialen Schicht dokumentieren, sind z.B. als empirische Basis für die Verifikation oder die Falsifikation einer bestimmten Hypothese geeignet. Das beantragte Einzelprojekt abstrahiert von diesen inhaltlichen Gesichtspunkten und betrachtet stattdessen die strukturellen Gemeinsamkeiten, die für die implizite Gegenstandskonstitution ganzer Theorieparadigmen kennzeichnend sind.

Innerhalb des Sonderforschungsbereichs besteht eine enge Verbindung zum Projekt A1, da die formalen Eigenschaften von Annotationssystemen durch die Einschränkung auf intendierte Modellklassen beeinflußt werden. Insbesondere vereint beide Projekte das Interesse an den Konsequenzen, die aus dem Grad der Strukturähnlichkeit der betrachteten Modelle zu endlichen Bäumen resultieren. Mit dem Projekt A5 besteht ein gemeinsames Interesse an Formalisierungen der HPSG. Hier soll das vorliegende Projekt durch seine allgemeinen Überlegungen die Grundlage für die speziellen linguistischen Untersuchungen im Projekt A5 legen. Der Zusammenhang mit dem Projekt B5 ist durch das gemeinsame Interesse an Grammatikformalismen gegeben. Die theoretischen Überlegungen zum Paradigmenwechsel in der Sprachwissenschaft sollen gemeinsam mit dem Projekt B13 durch exemplarische Überlegungen zur Ellipsenforschung praktisch nachgezeichnet werden. Die zu erwartenden theoretischen Ergebnisse des Projekts zu Fragen semistrukturierter Daten und Annotationsstandards sollten gerade auch in ihrem komplexitätstheoretischen Teil die Wahl der Annotation

sowohl des Projekts A1 als auch der phänomenorientierten Projekte, die Corpora erstellen, (B1, B3, B8, B9, B11) beeinflussen. Die bereits in der letzten Phase begonnene Zusammenarbeit mit den Projekten, die im weiteren Sinne Deixis erforschen (B8, B9), soll gerade im Bezug auf die Untersuchung lokaler Deixis im Warao fortgesetzt werden. Bei der Erforschung der amerindischen Sprache Warao werden wir mit dem Projekt B11, die beim Tibetischen vor ähnlichen Problemen der adäquaten Berücksichtigung von außersprachlichem Kontext wie Kultur, sozialen und ökologischen Bedingungen etc. stehen, Erfahrungen und Methoden austauschen und die Schemata von Annotationen aufeinander abstimmen.
Das vorliegende Projekt wird sich an dem gemeinsamen Workshop *Datentypenvergleich: Diachronie, Ontogenese, Typologie* der Projekte B1, B3, B6, B9 und A2 beteiligen und die formalen Grundlagen eines Datentypenvergleichs beisteuern.

Außerhalb des Sonderforschungsbereichs sind Kooperationen mit folgenden Fachkollegen vereinbart:

- Prof. Dr. Peter Aczel, Universität Manchester
- Prof. Dr. Dieter Heinen, Universität Caracas
- Prof. Dr. Aravind Joshi, Universität von Pennsylvania in Philadelphia
- Prof. Dr. Edward Keenan, Universität von Kalifornien in Los Angeles
- Prof. Dr. Lawrence Moss, Universität von Indiana in Bloomington
- Prof. Dr. Klaus Schulz, Universität München
- Prof. Dr. Thomas Schwentick, Universität Jena
- Peter Wittenburg, MPI Nijmegen

Natürlich wird eine enge Zusammenarbeit mit dem Teilprojekt C2 „Coalgebraische Modellierung von Hypertexten und Annotationsgraphen“ der distribuierten Forschergruppe *Texttechnologie* (Sprecher: Prof. Dr. Dieter Metzing, Bielefeld) stattfinden. Desgleichen ist beabsichtigt, die bestehenden guten Kontakte zu dem niederländischen NWO-Verbundprojekt *Typological Databases* weiter auszubauen.

Zitierte Literatur

Abiteboul, Serge (1997): „Querying Semi-Structured Data“, *in* Foto Afrati und Phokion Kolaitis, (Hrsg.), *Database Theory – Proceedings ICDT*, LNCS Nr. 1186, Springer-Verlag.

Abiteboul, Serge, Peter Buneman und Dan Suciu (2000): *Data on the Web*, Morgan Kaufmann.

Aczel, Peter (1997): „The Initial Algebra and the Final Coalgebra Perspectives“, *in* Helmut Schwichtenberg, (Hrsg.), *Logic of Computation*, Springer-Verlag, S. 1–33.

Barbosa, Pilar, Danny Fox, Paul Hagstrom, Mertha McGinnis und David Pesetsky, (Hrsg.) (1998): *Is the Best Good Enough? Optimality and Competition in Syntax*, MIT Press.

Barr, Michael und Charles Wells (1995): *Category Theory for Computing Science*, Prentice Hall.

Blackburn, Patrick und Edith Spaan (1993): „A Modal Perspective on the Computational Complexity of Attribute Value Grammar“, *Journal of Logic, Language, and Information* **2**, 129–169.

Bodlaender, Hans L. (1993): „A Tourist Guide through Treewidth“, *Acta Cybernetica* **11**, 1–23.

Börger, Egon, Erich Grädel und Yuri Gurevich (1997): *The Classical Decision Problem*, Springer-Verlag.

Chidlovskii, Boris (2000): „Using Regular Tree Automata as XML Schemas“, in *Proc. IEEE Advances in Digital Libraries Conference*.

Chomsky, Noam (1955/1975): *The Logical Structure of Linguistic Theory*, Plenum Press.

Chomsky, Noam (1981): *Lectures on Government and Binding*, Foris Publications, Dordrecht, Holland.

Chomsky, Noam (1986): *Barries*, MIT Press.

Chomsky, Noam (1995): *The Minimalist Program*, MIT Press.

Courcelle, Bruno (1997): „The Expression of Graph Properties and Graph Transformations in Monadic Second-Order Logic“, *in* Grzegorz Rozenberg, (Hrsg.), *Handbook of Graph Grammars and Computing by Graph Transformation*, World Scientific Publishing, S. 313–400.

Downey, Rodney Graham und Michael R. Fellows (1999): *Parameterized Complexity*, Springer-Verlag.

Ebbinghaus, Heinz-Dieter und Jörg Flum (1995): *Finite Model Theory*, Springer-Verlag.

Fagin, Ronald (1974): „Generalized First-Order Spectra and Polynomial-Time Recognizable Sets“, *in* Richard Karp, (Hrsg.), *Complexity of Computation*, SIAM-AMS Proceedings Nr. 7, S. 43–73.

Goguen, Joseph (1991): „Types as Theories“, *in* George M. Reed, A. William Roscoe und Ralph F. Wachter, (Hrsg.), *Topology and Category Theory in Computer Science*, Clarendon Press, S. 357–390.

Jacobs, Bart (1996): „Objects and Classes, Co-Algebraically“, *in* Burkhard Freitag, (Hrsg.), *Object-Orientation with Parallelism and Persistence*, Kluwer, S. 83–103.

King, Paul John, Kiril Ivanov Simov und Bjørn Aldag (1999): „The Complexity of Modellability in Finite and Computable Signatures of a Constraint Logic for Head-Driven Phrase Structure Grammar“, *Journal of Logic, Language and Information* **8**(1), 83–110.

McHugh, Jason und Jennifer Widom (1999): „Query Optimization for XML“, in *Pro-*

ceedings of IEEE Very Large Databases Conference, S. 315–326.

Meseguer, José und Joseph Goguen (1985): „Initiality, Induction, and Computability“, *in* Maurice Nivat und John Reynolds, (Hrsg.), *Algebraic Methods in Semantics*, Cambridge University Press, S. 459–541.

Murata, Makoto (1999): *Hedge Automata: A Formal Model for XML Schemata*, Technischer Bericht, Fuji Xerox Information Systems.

Murata, Makoto (2001): „Extended Path Expressions for XML“, *in* Peter Buneman, (Hrsg.), *Proceedings PODS 2001*, ACM.

Neven, Frank und Thomas Schwentick (2000): „Expressive and Efficient Pattern Languages for Tree-Structured Data“, *in* Bertram Ludäscher, (Hrsg.), *Proceedings PODS 2000*, ACM.

Pollard, Carl und Ivan A. Sag (1987): *Information Based Syntax and Semantics, Vol. 1: Fundamentals*, Lecture Notes Nr. 13, CSLI.

Pollard, Carl und Ivan A. Sag (1994): *Head-Driven Phrase Structure Grammar*, University of Chicago Press.

Postal, Paul (1964): *Constituent Structure*, Mouton.

Prince, Alan und Paul Smolensky (1993): *Optimality Theory: Constraint Interaction in Generative Grammar*, Technischer Bericht RuCCTS-TR 2, Rutgers University.

Putnam, Hilary (1961): „Some Issues in the Theory of Grammar“, in *Proceedings of Symposia in Applied Mathematics*, Band 12, S. 25–42.

Reichel, Horst (1995): „An Approach to Object Semantics Based on Terminal Coalgebras“, *Mathematical Structures in Computer Science* **5**, 129–152.

Rogers, James (1998): *A Descriptive Approach to Language-Theoretic Complexity*, CSLI Publications.

Rutten, Jan (2000): „Universal Coalgebra: A Theory of Systems“, *Theoretical Computer Science* **260**(1–2), 3–80.

Seese, Detlef (1991): „The Structure of the Models of Decidable Monadic Theories of Graphs“, *Annals of Pure and Applied Logic* **53**, 169–195.

Seese, Detlef (1992): „Interpretability and Tree Automata: A Simple Way to Solve Algorithms Problems on Graphs Closely Related to Trees“, *in* Maurice Nivat und Andreas Podelski, (Hrsg.), *Tree Automata and Languages*, North Holland, S. 83–114.

Thomas, Wolfgang (1997): „Languages, Automata, and Logic“, *in* Grzegorz Rozenberg und A. Salomaa, (Hrsg.), *Handbook of Formal Languages, Vol 3: Beyond Words*, Springer-Verlag, S. 389–455.

van Leeuwen, Jan (1990): „Graph Algorithms“, *in* Jan van Leeuwen, (Hrsg.), *Handbook of Theoretical Computer Science*, Band A: Algorithms and Complexity, Elsevier, S. 525–631.

von Stechow, Arnim und Wolfgang Sternefeld (1988): *Bausteine syntaktischen Wissens*, Westdeutscher Verlag.

3.7 Ergänzungsausstattung für das Teilprojekt

Es bedeuten:
PK: Personalbedarf und –kosten (Begründung vgl. 3.7.1)
SV: Sächliche Verwaltungsausgaben (Begründung vgl. 3.7.2)
I: Investitionen (Geräte über DM 20.000,– brutto; Begründung vgl. 3.7.3)

	Bewilligung 2001			2002			2003			2004		
PK	Verg.-Gr.	An-zahl	Betrag in DM	Verg.-Gr.	An-zahl	Betrag in DM	Verg.-Gr.	An-zahl	Betrag in DM	Verg.-Gr.	An-zahl	Betrag in DM
	BAT IIa	1	106.800	BAT IIa	1	106.800	BAT IIa	1	106.800	BAT IIa	1	106.800
	BAT IIa/2	1	48.000	Wiss.HK	1	33.600	Wiss.HK	1	33.600	Wiss.HK	1	33.600
				Stud.HK	0,5	10.200	Stud.HK	0,5	10.200	Stud.HK	0,5	10.200
	zusammen		154.800	zusammen		150.600	zusammen		150.600	zusammen		150.600
SV				Kostenkategorie oder Kennziffer		Betrag in DM	Kostenkategorie oder Kennziffer		Betrag in DM	Kostenkategorie oder Kennziffer		Betrag in DM
				Verbrauchsmittel		1.000	Verbrauchsmittel		1.000	Verbrauchsmittel		1.000
				zusammen		1.000	zusammen		1.000	zusammen		1.000
I				Mittel für Invest. insgesamt: 0			Mittel für Invest. insgesamt: 0			Mittel für Invest. insgesamt: 0		

3.7.1 Begründung des Personalbedarfs

	Name, akad. Grad, Dienststellung	engeres Fach des Mitarbeiters	Institut der Hochschule oder der außeruniv. Einrichtung	Mitarbeit im Teil-projekt in Stunden pro Woche	auf d. Stelle im SFB tätig seit	beantr. Einstu-fung in BAT…
Grund-ausstattung						
3.7.1.1 wissenschaftl. Mitarbeiter	1. U. Mönnich, Prof. Dr.	Theor. Com-puterlinguistik, Vergleichende Sprachwiss.	SfS	8		
(einschl. Hilfskräfte)	2. F. Morawietz, MA, Wiss. Ang.	Formale Linguistik	SfS	8		
	3. E. Weißhar, Dr., Akad. Oberrat	Ethnoling.	Seminar für Vergl. Sprachwiss.	5		
Ergänzungs-ausstattung						
3.7.1.3 wissenschaftl. Mitarbeiter	4. St. Kepser, Dr.	Theor. Com-puterlinguistik	SFB 441	38,5	1.3. 1999	IIa
(einschl. Hilfskräfte)	5. St. Herrmann, MA	Ethnoling.	SFB 441	20	1.3. 1999	Wiss. HK.
	X 6. N.N.	Computer-linguistik		10		Stud. HK.
3.7.1.4 nichtwissen-schaftliche Mitarbeiter						

(Stellen, für die Mittel neu beantragt werden, sind mit **X** gekennzeichnet)

1. Der Antragsteller wird das Teilprojekt leiten und die damit verbundenen Forschungsarbeiten betreuen. Neben seiner Lehrtätigkeit und der Anleitung von Abschlußarbeiten wird er sich hauptsächlich auf den Teil des Arbeitsprogramms konzentrieren, in dem die Berechenbarkeit von Datenstrukturen untersucht wird.

2. Der Schwerpunkt der Interessen von Frank Morawietz liegt seit seiner Promotion über Baumbeschreibungssprachen auf dem Gebiet der Formalisierung generierender linguistischer Theorien. F. Morawietz kann daher nicht nur als einschlägig ausgewiesen gelten für die Mitarbeit am geplanten Einzelprojekt, sondern sein besonderes Interessengebiet, das in mehreren Publikationen zu diesem Thema dokumentiert ist, stellt eine wünschenswerte Ergänzung der von St. Kepser eingebrachten Erfahrung dar, da in der Verteilung der Hauptarbeitsgebiete dieser vorgesehenen wissenschaftlichen Mitarbeiter die Entwicklung von dem Konzept abstrakter Datenstrukturen zu der Auffassung von Theorien als Datentypen in angemessener Weise repräsentiert ist.

3. Bei der Bewertung der Daten des Warao wird Frau Herrmann von Dr. Emmerich Weißhar unterstützt werden, dem Ethnolinguisten am hiesigen Seminar für Vergleichende Sprachwissenschaft, der als Berater für das Projekt hat gewonnen werden können, und der seit vielen Jahren über intensive eigene Erfahrungen im mittel- und südamerikanischen Sprachraum verfügt.

4. Dr. Stephan Kepser wird wie schon in der zurückliegenden Phase hauptsächlich die Forschungsarbeit zur algebraischen und coalgebraischen Modellierung linguistischer Datenstrukturen und zu deren Berechenbarkeit leisten.

5. Um die in der zurückliegenden Phase gesammelten und zum Teil aufbereiteten Daten nicht verloren gehen zu lassen, wird Stefanie Herrmann ihre Analyse der Waraodaten zu Ende führen und diese in elektronischer Form aufarbeiten und in Datenbanken und Corpora ablegen, um sie, sobald die Sprachdokumentationssysteme vom MPI in Nijmegen zur Verfügung stehen, in diese zu integrieren und damit für die interessierte wissenschaftliche Öffentlichkeit verfügbar zu machen.

6. Neben den üblichen wissenschaftlichen und organisatorischen Hilfstätigkeiten und der Vermittlung zu anderen Projekten soll die Hilfskraft insbesondere bei der Recherche und Auflistung der Art der verwendeten linguistischen Annotationen in Dokumenten in Open Language Archives (OLAC) unterstützend tätig werden.

3.7.2 Aufgliederung und Begründung der sächlichen Verwaltungsausgaben (nach Haushaltsjahren)

	2002	2003	2004
Für Sächliche Verwaltungsausgaben – stehen als Grundausstattung voraussichtlich zur Verfügung	19.000,–	19.000,–	19.000,–
– werden als Ergänzungsausstattung beantragt (entspricht den Summen „Sächliche Verwaltungsausgaben“ in Übersicht 3.7)	1.000,–	1.000,–	1.000,–

Grundausstattung

Aus der Grundausstattung des Antragstellers wurden eine Workstation und ein Arbeitsplatz-PC im Wert von insgesamt 14.000,– DM angeschafft. Ebenfalls aus der Grundausstattung stammen Bücher im Wert von zusammen 3.000,– DM. Der Antragsteller wird jährlich 2.000,– DM aus seinen Sachmitteln für das Teilprojekt verwenden.

Ergänzungsausstattung

Verbrauchsmittel **(522)** Büromaterialien und EDV-Zubehör.

Reisemittel, Gastmittel und Mittel für Kopien werden für den gesamten SFB zentral vom Projekt Z beantragt.

3.7.3 Investitionen

(Geräte über DM 20.000 brutto und Fahrzeuge)

Es werden keine Investitionsmittel beantragt.