3.1 Allgemeine Angaben zum Teilprojekt A1

3.11 Thema

Repräsentation und Erschließung linguistischer Daten

3.12 Fachgebiet und Arbeitsrichtung

Syntax des Deutschen, Computerlinguistik, Korpuslinguistik.

3.13 Leiter

Name:	Hinrichs, Erhard, Prof. Dr.
E-Mail:	eh@sfs.nphil.uni-tuebingen.de

3.2 Zusammenfassung

Das Verhältnis von Empirie und Theorie in der Grammatikforschung als zentraler Forschungsgegenstand des beantragten SFBs soll im Teilprojekt A1 am Gegenstand von elektronisch verfügbaren Textkorpora untersucht werden.
Die Verfügbarkeit großer, elektronisch gespeicherter Textkorpora eröffnet eine neue Datenquelle, die sich zu einem wichtigen Instrument der linguistischen Theoriebildung zu entwickeln verspricht. Die Nützlichkeit in Korpora gesammelter Datenmengen hängt vor allem von der Art ihrer Aufbereitung ab. Hierbei hat nach der anfänglichen Konzentration auf kleine Annotationseinheiten (Wortartentagging) zunehmend die Betrachtung von komplexen syntaktischen Strukturen an Bedeutung gewonnen. Allerdings sind elaborierte Darstellungen von syntaktischen Phänomenen stark analyseabhängig, so daß ein einmal annotiertes Korpus nur selten in unveränderter Form einer allgemeinen Verwendung zugeführt werden kann. Dadurch sind syntaktisch annotierte Korpora trotz des in der Regel hohen Zeit- und Arbeitsaufwandes bei der Annotierung oft nur beschränkt wiederverwertbar. Für das beantragte Projekt ergibt sich deshalb eine zweifache Zielsetzung: Zum einen werden geeignete theorieneutrale *Repräsentationsformen* für die Daten definiert, zum anderen wird die *Erschließbarkeit* der Daten untersucht. Diese beiden Ziele sollen in drei Bereichen des Projekts verfolgt werden. Die drei Bereiche gruppieren sich um die Bereitstellung komplexer automatischer Annotationen, die Abfrage in komplex annotierten Korpora und die weitere linguistische Aufarbeitung komplex annotierter Korpora.
Automatische inkrementelle Annotation: Für verschiedene Annotationsebenen (vor allem Chunks und Sätze) werden Repräsentationssysteme zugrunde gelegt, die eine informationsreiche, theorieneutrale Annotation ermöglichen. Es werden Verfahren untersucht, die eine inkrementelle Abbildung von einer Annotationsebene in die nächsthöhere erlauben. Es geht also sowohl um die Wahl geeigneter Repräsentationssysteme als auch um die automatische Erschließung einer Annotationsebene aus der darunterliegenden durch den Einsatz maschineller Lernverfahren.
Annotationserschließung: Syntaktische Annotationssysteme sind in der Regel baumbasiert. Zur systematischen Erschließung der Daten in syntaktisch annotierten Korpora wird eine Querysprache benötigt, die der Baumannotation Rechnung trägt. Hierzu wird eine entsprechende Baumdeskriptionssprache definiert. In diesem Bereich ist die Deskriptionssprache als

Repräsentation einer Query zentral, die dann zur Erschließung von annotierten Daten aus einem Korpus verwendet werden kann.
Theoriegeleitete Annotationsveränderung: Es wird ein allgemeines Verfahren definiert, das den Übergang zwischen quantitativen Daten und qualitativen Daten ermöglicht. Ziel ist es, ein vorhandenes Korpus in ein Korpus zu überführen, das genau die Information beinhaltet, die von einer bestimmten linguistischen Theorie angenommen wird. Die Datenrepräsentation betrifft hier primär den Status theorieorientierter Annotationen, die notwendigerweise immer partiell sind. Hinsichtlich der Datenerschließung wird die Abbildung von robusten, vollständigen Annotationen zu problemorientierten, detaillierten aber partiellen Annotationen untersucht.
In allen drei Bereichen steht die Untersuchung und Entwicklung von generellen Verfahren und Repräsentationssystemen für die Erstellung und Erschließung linguistisch annotierter Korpora im Vordergrund. Gleichzeitig sollen die Ergebnisse dieser Untersuchungen den Teilprojekten des SFBs, bei denen die datenorientierte Untersuchung linguistischer Einzelphänomene im Vordergrund steht, mit dem Ziel zur Verfügung gestellt werden, die Validität der entwickelten Verfahren und Repräsentationssysteme empirisch zu überprüfen.

3.3 Stand der Forschung

Automatische inkrementelle Annotation

Korpora können auf verschiedenen Ebenen annotiert werden. Hierbei gilt es vor allem, die folgenden Ebenen zu unterscheiden:

- bloßer, unannotierter Text
- Text mit Wortarten-Tagging
- Zusammenfassen von Wortgruppen zu (Teil-)Phrasen (Chunking)
- syntaktische Struktur von Sätzen (meist Bäume)
- Angabe der Argumentstruktur
- ...
- vollständige semantische Analyse des Texts

Anhand der verschiedenen Annotationsebenen soll der Stand der Forschung dargestellt werden.

Die Ebene der Wortartenbestimmung (*part-of-speech tagging*) ist in jüngeren Korpusarbeiten die bestuntersuchte. Hier haben sich bereits Standards herausgebildet. Mit dem *Stuttgart-Tübingen Tag-Set* (STTS), ((Thielen und Schiller 1994), (Schiller, Teufel und Thielen 1995)) ist für das Deutsche eine rein auf syntaktischer Distribution beruhende Repräsentationsform vorgeschlagen worden, die in das EAGLES-Projekt Eingang gefunden hat (EAGLES 1996b).

Das STTS ist dabei so angelegt, daß die Tags durch beliebige weitere Informationen angereichert werden können. Damit ist auf Wortartenebene eine geeignete theorieneutrale Repräsentation gegeben, die für problemorientierte oder theoriegeleitete Annotationsveränderung offen ist.

Als nächsthöhere Annotationsebene soll der wortübergreifende Bereich angesprochen werden. Abney (1991) definiert einen Chunk-Begriff, der ein Inhaltswort zusammen mit syntaktisch assoziierten Funktionswörtern zu einer Einheit zusammenfaßt. Von ihrer Konzeption her sind Chunks nichtrekursiv, d.h., ein Chunk besteht aus Wörtern, nicht aber aus Chunks. (Abney 1992) zeigt die linguistische Relevanz von Chunks. So kann im Englischen innerhalb eines nominalen Chunks kein weiterer Chunk auftreten (1a). Die Übersetzung des englischen Beispiels in (1b) zeigt, daß diese Beschränkung nicht in gleicher Weise für das Deutsche gilt.

(1) a. * the proud [$_{PP}$ of his son] man

b. der [$_{PP}$ auf seinen Sohn] stolze Mann

Trotz dieses Unterschieds ist unbestritten, daß auch im Deutschen nichtrekursiven Strukturen beim Aufbau der syntaktischen Gesamtstruktur eine wichtige Rolle zukommt.

Nach der Wort- und der Chunkebene kann nun die höchste im Projekt zu behandelnde syntaktische Repräsentationsebene betrachtet werden, die Satzebene. Soll ganzen Sätzen eine syntaktische Struktur zugeordnet werden, so ist in naher Zukunft kein Standard in Sicht, da die notwendigen Strukturannahmen sowie die Annotationslabels theorieabhängig sind (EAGLES 1996c, EAGLES 1996a). Das Grundproblem liegt hierbei darin, daß die syntaktische Struktur eine Verbindung herstellen soll zwischen der Argumentselektion und der Wortstellung. Verbzweit- und Kohärenzphänomene (Bech 1955), Stellungsfreiheiten im Mittelfeld (Eisenberg 1989, Kapitel 12), Extraposition, sowie die Regularitäten der Vorfeldbesetzung führen jedoch zu einer großen Differenz zwischen angenommener Argumentstruktur und beobachteter Wortstellung. Die daraus resultierenden strukturellen Operationen (Bewegungen verschiedener Art im GB-Rahmen; bestimmte Merkmalsmechanismen, Argumentanhebung und lexikalische Regeln in der HPSG) führen zu sehr unterschiedlichen Analysen, die größtenteils nicht miteinander kompatibel sind.
Das topologische Feldermodell des Deutschen, wie es Engel (1970) und Höhle (1986) zusammenfassen, hat sich in der Praxis der Korpusannotation als robuste und theorieneutrale Repräsentation erwiesen, mittels derer im laufenden Verbundprojekt *Verbmobil-II* ein Korpus von 30.000 Sätzen annotiert wird. Das Feldermodell geht primär von der tatsächlichen Wortstellung aus. Die Argumentstruktur findet sich lediglich in bestimmten Besetzungsbeschränkungen für einzelne Felder wieder. Alle syntaktischen Theorien des Deutschen müssen die grundlegenden topologischen Phänomene beschreiben. Die topologischen Felder tauchen jedoch meist nicht explizit in den Theorien auf ((Kathol 1995) ist die Ausnahme), da sie keinerlei explikativen Charakter haben. Allerdings weist Höhle (1986) auf einige Probleme des Feldermodells hin, die in der Forschung innerhalb der Feldertheorie noch nicht gelöst worden sind.

Parallel zur Verbmobil-Baumbank wird im Rahmen des Teilprojekts C3 des SFB 378 (Uszkoreit/Smolka, NEGRA, Nebenläufige grammatische Verarbeitung) ein zweites syntaktisch annotiertes Korpus entwickelt. Die beiden angesprochenen Baumbanken befinden sich noch in der Entwicklungsphase und werden unter großem personellen und zeitlichen Aufwand erstellt. Bei manuell annotierten Korpora besteht jedoch immer das Problem, daß selbst bei Anwendung größter Sorgfalt Inkonsistenzen und Fehler auftreten. Dieses Problem läßt sich nur beheben, indem automatische Annotationsverfahren (*Parser*) verwendet werden. Charniak (1996) vergleicht verschiedene Parser und Grammatiken mit einer automatisch aus dem Wall Street Journal extrahierten Grammatik. Letztere liefert bei dem Vergleich die besten Ergebnisse, aber selbst sie kann bei längeren Sätzen bis zu 40 Wörtern nur einen Korrektheitsgrad von 78,8 % erreichen. Sätze, die über eine Länge von 40 Wörtern hinausgehen, wurden bei diesem Experiment nicht berücksichtigt, da sie die z.Z. verwendeten Parser oft vor unüberwindbare Probleme stellen. Außerdem wurden nur Brackets (d.h. nur Konstituentenmarkierungen) verglichen, keine Labels, was bei weitem schwieriger wäre.

Eine weitere Möglichkeit, syntaktische Annotation automatisch durchführen zu lassen, bieten Lernverfahren. Linguistische Probleme unterscheiden sich nicht grundsätzlich von Problemen in anderen Bereichen, in denen heutzutage schon maschinelle Lernverfahren angewandt werden. Auch (computer-)linguistische Aufgabenstellungen lassen sich als Segmentierungs- und Klassifizierungsprobleme formulieren. In der Spracherkennung muß z.B. eine Segmentierung des Sprachsignals in Phone und eine Klassifizierung dieser Phone in Phoneme geleistet werden. Daher lassen sich auch in diesem Gebiet generelle maschinelle Lernverfahren verwenden. Von den potentiell anwendbaren Lernverfahren sollen kurz diejenigen beschrieben werden, die direkt für die Arbeit des Projekts einschlägig sind. Da auch für unrestringierte Texte allgemeines linguistisches Wissen über die Gestalt der möglichen syntaktischen Annotation vorliegt, ist es wichtig, Verfahren zu wählen, die es erlauben, das vorhandene linguistische Wissen mit zu berücksichtigen. Die Lernverfahren werden deshalb hinsichtlich des Einflusses , den der Anwender auf sie nehmen kann, aufgelistet.
Bei rein statistischen Verfahren kann linguistisches Wissen lediglich durch die Wahl des Annotationsschemas explizit miteingebracht werden. Das im Bereich Annotation wichtigste dieser Verfahren sind die Hidden-Markov-Modelle (HMM) (Church und Mercer 1993). HMMs sind probabilistische Übergangsnetzwerke, in denen aus den Übergangswahrscheinlichkeiten der wahrscheinlichste Pfad für eine Eingabe ermittelt wird. Beim Tagging wird z.B. das wahrscheinlichste Tag unter Berücksichtigung eines Kontextes von 1 - 2 vorausgehenden Wörtern ermittelt. Das Lernen besteht hier aus der automatischen Ermittlung der Wahrscheinlichkeiten der Übergänge aus einem Trainingskorpus. Da jedoch nicht alle Bi- oder Trigramme im Korpus enthalten sind, müssen die fehlenden abgeschätzt werden. Bei (Cutting, Kupiec, Pederson und Sibun 1992) werden Hidden-Markov-Modelle zum Tagging eingesetzt. (Brants und Skut 1998) verwenden solche Modelle, um die manuelle syntaktische Annotation eines Korpus zu unterstützen.
Stochastic Tree-Substitution Grammars (STSG) wie DOP (*data-oriented language processing*, Bod und Scha (1996)) stellen eine Generalisierung des n-Gramm-Begriffs auf syntaktische Teilbäume dar. Eine STSG wird aus einer Trainingsbaumbank gewonnen, in dem Teilbäume der Baumbank mit ihrer Häufigkeit extrahiert werden. Stochastische kontextfreie Grammatiken sind ein Spezialfall von STSGs, die nur Teilbäume mit Tiefe 1 enthalten. Im allgemeinen Fall muß bei STSGs zwischen Derivationsgeschichte und syntaktischer Struktur (Parse) unterschieden werden, wobei die Frage nach der wahrscheinlichsten Derivation entscheidbar, die nach dem wahrscheinlichsten Parse jedoch unentscheidbar ist. Die besten Ergebnisse werden in diesem Paradigma erzielt, wenn möglichst tiefe Teilbäume in die Grammatik mit aufgenommen werden und ein möglichst wahrscheinlicher Parse gesucht wird.

Eine andere Form der Extension von HMMs auf Baumstrukturen stellen Entscheidungsbaum-Modelle dar (Magerman 1995). Hierbei entspricht die Tiefe des Entscheidungsbaums den n-Grammen der HMMs. Auch bei Entscheidungsbäumen werden die Übergangswahrscheinlichkeiten zwischen Knoten aus einem Trainingskorpus gelernt. Entscheidungsbaumverfahren erlauben dem Anwender jedoch, konkretes linguistisches Wissen in die Form der möglichen Entscheidungen einzubringen. Magerman (1995) zeigt, daß ein solches Verfahren für syntaktische Annotation von Korpora bessere Ergebnisse liefert als grammatikbasierte Parser.
Memory-based learning (auch *similarity-based, case-based, exemplar-based* oder *nearest-neighbor* genannt) (Daelemans 1995, Argamon, Dagan und Krymolowski 1998) ist ein induktiver Lernansatz, der durch überwachtes Lernen aus einem Beispielkorpus eine Similaritätsmatrix berechnet. Wird dem System ein neuer Fall präsentiert, kann durch die Similaritätsmatrix das ähnlichste Beispiel ermittelt werden. Linguistisches Wissen kann hierbei direkt in der Similaritätsfunktion enkodiert werden. Eine besondere Eigenschaft des Ansatzes liegt darin, daß die Trainingsinstanzen unverändert im Speicher bleiben. Dadurch wird der Tatsache Rechnung getragen, daß in unbeschränkten Texten viele Subregularitäten und Unregelmäßigkeiten der Sprache auftreten (Daelemans 1998), deren Mißachtung das Ergebnis des Lernverfahrens negativ beeinflussen.
Das *transformation-based error-driven learning* (Brill 1993b) ist ein schwach-statistisches regelbasiertes Lernverfahren, in das viel linguistisches Wissen direkt einfließen kann. Deshalb sind auch bereits mit einem verhältnismäßig kleinen Trainingskorpus Ergebnisse möglich, die mit anderen Ansätzen vergleichbar sind. Das Lernverfahren geht von einem Korpus aus, das trivial annotiert ist (z.B. mit zufälligen Annotationen). Durch den Vergleich mit dem Trainingskorpus werden iterativ Transformationen ermittelt, die die Annotationen so verändern, daß sie dem Trainingskorpus möglichst ähnlich werden. Aus den Transformationshypothesen wird diejenige ausgewählt, die den größten positiven Effekt hat, und auf das Korpus angewendet. Das Verfahren wurde im Bereich Tagging (Brill 1993b), Syntax-Bracketing (Brill 1993a), Disambiguierung von Präpositionalphrasenanbindung (Brill und Resnik 1994) und Text Chunking (Ramshaw und Marcus 1995) angewandt.
Ein rein regelbasiertes Lernverfahren ist mit dem *explanation-based learning* (Samuelsson 1994, Neumann 1997) gegeben. Bei diesem Verfahren wird ausgehend von einer großen Menge von Regeln und einem Trainingskorpus auf eine kleinere, für das Korpus ausreichende Menge von Regeln geschlossen. Es liegt eine eindeutige Dominanz auf Seite des linguistischen Wissens vor, da das rein deduktive Verfahren nicht in der Lage ist, zusätzliche Regeln oder Regelveränderungen aus dem Trainingskorpus zu erschließen.

Mit zunehmender Einflußmöglichkeit des Anwenders auf das Lernverfahren steigt zwar die Korrektheit bei vollständig analysierten Strukturen, allerdings sinkt die Robustheit des Verfahrens. Das Projekt wird sich deshalb auf Lernverfahren des mittleren Bereichs beschränken.

Zusammenfassend ergibt sich die folgende Problematik: Die bislang für das Deutsche existierenden Baumbanken sind manuell und unter großem zeitlichen und finanziellen Aufwand erstellt worden, so daß dringend nach automatischen Alternativen Ausschau gehalten werden muß. Automatische Verfahren sind andererseits bislang nur für einzelne Annotationsebenen erprobt worden. Dabei hat sich gezeigt, daß für die Ebenen des Taggings und Chunkings gute Ergebnisse erzielt werden können. Bei automatischer, vollständiger syntaktischer Analyse in einem Schritt konnten jedoch bislang keine befriedigenden Ergebnisse erzielt werden. Die im Projekt angestrebte inkrementelle Architektur reduziert das Problem der vollständigen syntaktischen Annotation in zweierlei Hinsicht. Zum einen erlaubt das verwendete Annotationsformat gezielt ausgezeichnete Formen der Unterspezifikation, zum anderen ist durch die Chunkebene eine zuverlässige syntaktische Teilannotation erreicht, auf der das Lernverfahren aufbaut.

Annotationserschließung

Bei vorhandenen Korpora stellt sich häufig das Problem, daß die Annotationen des Korpus zwar möglicherweise sehr reichhaltige und interessante strukturelle Information enthalten, daß jedoch die Werkzeuge zur Erschließung der Annotationen fehlen und daher die in den Annotationen enthaltene Information nicht zugänglich ist.
Ein bestehendes Anfragesystem für Korpora ist das in Stuttgart entwickelte Xkwic ((Christ 1994) und (Christ und Schulze 1996)), ein System, das Anfragen mit regulären Ausdrücken erlaubt. Diese regulären Ausdrücke können kombiniert werden mit verschiedenen Anweisungen, die teilweise die zu suchenden Teilausdrücke weiter einschränken und teilweise Anweisungen an das Suchprogramm sind. Xkwic bezieht sich auf die lineare Abfolge einzelner Wörter oder Teilausdrücke, das heißt eine Anfrage mit mehreren Teilausdrücken, deren lineare Präzedenz nicht bekannt ist bzw. offen gelassen werden soll, ist nur durch eine entsprechende Disjunktion möglich. Reguläre Ausdrücke reichen nicht aus, um Anfragen zu stellen, bei denen die lineare Abfolge der einzelnen Teilstrukturen nicht völlig festgelegt ist. Mit einem System wie Xkwic ist eine Suche nach Teilstrukturen höchstens dann möglich, wenn eine syntaktische Annotation nicht als explizite Baumstruktur über einem Satz kodiert sondern durch eine entsprechende Klammerung der Konstituenten dargestellt ist.
Für die Suche in syntaktisch annotierten Korpora ist von Denys Duchier im NEGRA-Projekt ein Anfragesystem entwickelt worden. Duchiers Anfragesprache erlaubt es, sich auf strukturelle Relationen wie unmittelbare Dominanz (direkte Nachfolger-Beziehung), transitive bzw. reflexive transitive Hülle von unmittelbarer Dominanz und auf lineare Präzedenz zu beziehen. Von ihrer Ausdrucksstärke her bietet diese Querysprache die Möglichkeiten einer Logik erster Ordnung mit Quantoren. Anfragen werden jedoch nicht als Ausdrücke in einer Logik formuliert, sondern als Programm.
Das von Duchier entwickelte Anfragesystem ist weitgehend durch praktische bzw. anwendungsorientierte und weniger durch mathematische Gesichtspunkte motiviert. Zu der von ihm verwendeten Anfragesprache gibt es keine denotationelle Semantik, und zwischen einem Ausdruck in der Anfragesprache und den gesuchten syntaktischen Strukturen gibt es dementsprechend keine direkte modelltheoretische Verbindung. Damit besteht der Zusammenhang zwischen der Query und dem Suchergebnis letztendlich nur auf intuitiver Ebene und ist durch die konkrete Implementierung des Querysystems bestimmt. Es handelt sich also bei dieser Anfragesprache nicht um eine Baumlogik im eigentlichen Sinne. Daneben sind bisher keine weiteren Entwicklungen von Anfragesprachen bekannt, die einen Bezug auf strukturelle Gegebenheiten erlauben, das heißt, bei denen es sich direkt um Baumbeschreibungssprachen handelt.
Arbeiten zu Baumlogiken gibt es in verschiedenen Bereichen, allerdings nicht im Zusammenhang mit Anfragesprachen für Textkorpora. In den meisten Fällen geht es um modelltheoretische Formalisierungen von Grammatiken, beispielsweise im Rahmen der Government

and Binding Theorie (GB), der Lexical Functional Grammar (LFG) oder der Tree Adjoining Grammar (TAG). Dies ist selbstverständlich ein völlig anderes Ziel als das einer Anfragesprache für Korpora, denn eine Grammatik hat den theoretischen Anspruch, (idealerweise) genau eine natürliche Sprache zu beschreiben, das heißt, eine geeignete Logik muß es in diesem Fall auch erlauben, allgemeine grammatische Prinzipien auszudrücken. Eine Querysprache dagegen soll dazu dienen, Anfragen zu stellen, also aus linguistischer Sicht interessante Aspekte von Daten zu formulieren. Dabei kann (in der Regel) davon ausgegangen werden, daß das untersuchte Korpus nur grammatikalisch korrekte Strukturen enthält. Die Anfragesprache muß demnach nicht die Formalisierung allgemeiner Regeln oder Prinzipien erlauben. Trotz dieses Unterschiedes sind die bisher vorgeschlagenen Baumlogiken sicher auch für eine Anfragesprache zur Annotationserschließung interessant. Im folgenden werden einige der hier zu berücksichtigenden Arbeiten genannt.
Zu erwähnen ist zunächst die quantorenfreie Logik erster Ordnung, die in (Rogers 1994) und (Rogers und Vijay-Shanker 1994) im Zusammenhang mit TAGs vorgeschlagen wurde, ebenso Modallogiken für Bäume wie z.B. in (Blackburn und Meyer-Viol 1994) behandelt. Logiken höherer Ordnung, die als Baumbeschreibungen vorgeschlagen wurden, sind z.B. $L^2_{K,P}$, eine in (Rogers 1994, Rogers 1996) eingeführte monadische Prädikatenlogik zweiter Ordnung (*monadic second order (MSO) logic*), die Rogers zu einer weitgehenden Formalisierung von GB verwendet. MSO Logiken sind entscheidbar und Endlichkeit kann in einer solchen Logik ausgedrückt werden. Im Rahmen des Projektes A8 (Mönnich) des SFB 340 sind ebenfalls Arbeiten zu MSO-Logik entstanden, beispielsweise (Morawietz und Cornell 1997) zum Zusammenhang zwischen MSO-Constraints und Baumautomaten. Eine weitere interessante Baumlogik höhere Ordnung sind die in (Niehren, Pinkal und Ruhrberg 1997a, Niehren, Pinkal und Ruhrberg 1997b) eingeführten *context constraints over finite trees*, die es erlauben, durch eine Relation *equality up-to* Parallelitäten zwischen Strukturen auszudrücken.
Diese verschiedenen Baumlogiken sollen in Betracht gezogen werden bei der Frage, welche Logik für die gesuchte Anfragesprache verwendet werden soll, bzw. wie Baumbeschreibungen aussehen müssen, die der Erschließung syntaktischer Annotationen dienen sollen.

Theoriegeleitete Annotationsveränderung

Das beantragte Projekt wirft ein neues Licht auf die Beziehung zwischen quantitativen und qualitativen Daten. Bisherige Ansätze konzentrieren sich in der Regel auf höchstens einen dieser Datentypen. Dies soll im folgenden an einigen Beispielen illustriert werden.
NEGRA: Im NEGRA-Projekt des SFB 378 entsteht ein großes, syntaktisch annotiertes Korpus. Bei der Erstellung von Baumbanken liegt der Schwerpunkt auf der Robustheit des Annotationssystems, nicht auf der linguistischen Adäquatheit einzelner Ananlysen. Das Annotationssystem strebt daher eine möglichst theorieneutrale Darstellungsform an, wobei das Korpus primär im Projekt selbst zur Entwicklung einer performanzorientierten Grammatiktheorie genutzt wird. Für das Projekt mußte das genannte Korpus speziell erstellt werden. Hierzu wurde ein neues interaktives Mensch-Maschine-Annotationsparadigma verwendet. Das Korpus enthält syntaktisch annotierte Sätze, wobei das Annotationsformat auch überkreuzende Äste bei Bäumen zuläßt. Es wäre nun wünschenswert, Korpora, die gemäß anderer Repräsentationssysteme annotiert wurden, für die Forschung in NEGRA verfügbar zu machen. Bislang muß die Veränderung von Annotationen jedoch auf inzidentiellen Über-

einstimmungen zwischen den Systemen und ad hoc-Abbildungsregeln beruhen.

ParGram: Das ParGram-System (Xerox Parc, Xerox Research Center Europe, IMS Stuttgart) ist eine LFG-Grammatik mit großer Abdeckung. Durch ihre Abdeckung und die relativ schnelle Laufzeit kann ein syntaktisch annotiertes Korpus durch Parsen erstellt werden. Das zugrundeliegende Annotationssystem sind Paare von C- und F-Strukturen. Das System ist relativ robust und die Analyse ist weitgehend linguistisch motiviert. Dennoch müssen sowohl hinsichtlich der Robustheit als auch hinsichtlich der Analyse Kompromisse geschlossen werden. So ist nicht für jeden Satz garantiert, daß eine (und möglichst genau eine) Analyse ausgegeben wird. Was die linguistische Adäquatheit betrifft, so können keine Analysen aufgenommen werden, die zu schlechten Parsezeiten führen.

ALE: Implementationsplattformen wie ALE (Carpenter und Penn 1996) sind dafür konzipiert, kleine, linguistisch interessante Fragmente zu implementieren. Diese Implementierungen sollen der Überprüfung und Weiterentwicklung linguistischer Theorien dienen. Bei Implementierungsplattformen geht es also um rein qualitative Daten. Die implementierten Theorien müssen allerdings in gewisser Hinsicht total sein, d.h., sie müssen eine Analyse für gesamte Sätze bereitstellen. Linguistische Theorien sind aber in der Regel nur an einzelnen Aspekten von Sätzen interessiert; ,,Fragmentgrammatiken" im Montagueschen Sinne sind nur ein Spezialfall, der in der praktischen linguistischen Arbeit selten auftritt. Durch die starke Spezialisierung linguistischer Einzeltheorien ist letztendlich gewährleistet, daß theoretische Erkenntnisse eines Ansatzes auch in anderen Ansätzen wiederverwendet werden können. Solche spezialisierten Analysen sind sehr detailliert hinsichtlich bestimmter Einzelaspekte von Sätzen, jedoch völlig oder teilweise neutral bezüglich den meisten anderen Aspekten. Diese Theorien sind also detailiert und partiell. Für sie können Implementierungsplattformen nur in den seltensten Fällen sinnvoll genutzt werden.

Eine theorieorientierte Annotation von Daten kann also im allgemeinen nur partiell sein. Die Repräsentation partieller Strukturen als Annotationsform ist jedoch bislang nicht untersucht. Sie findet allerdings eine Entsprechung im Bereich der Unterspezifiziertheit, wie sie in der computerlinguistischen Semantik bereits etabliert ist. (Reyle 1993) definiert unterspezifizierte Repräsentationen, die die Menge der disambiguierten Repräsentationen denotieren. In (Bos 1995) wird eine allgemeine Methode vorgestellt, für beliebige semantische Objektsprachen unterspezifizierte Entsprechungen zu definieren.

Auf der Ebene der Syntax wurde in (Pollard und Sag 1987) der Versuch gemacht, unterspezifizierte linguistische Objekte in die Syntaxforschung zu integrieren. Durch eine unsaubere Trennung zwischen linguistischen Objekten und Beschreibungen ergab sich jedoch in diesem Ansatz kein in irgendeiner Weise kohärentes Bild (vergleiche die Kritik in (King 1989, Carpenter 1992)).

Als exemplarischer Phänomenbereich soll im Projektbereich "Theoriegeleitete Annotationsveränderung" die Verbsyntax des Deutschen betrachtet werden. Es handelt sich hierbei um einen gut untersuchten empirischen Bereich, dem sich viele Einzeltheorien zugewendet haben (Bech 1955, Evers 1975, Haegeman und Riemsdijk 1986, den Besten und Edmondson 1983, von Stechow und Sternefeld 1988, Kroch und Santorini 1991, Haider 1993, Kiss 1993, Kathol 1995, ...), und der im SFB 1711 eine zentrale Rolle in den germanistisch orientierten Projekten B3 (Ehrich/Reis) und B10 (von Stechow) spielt.

3.4 Eigene Vorarbeiten

Das beantragte Forschungsvorhaben beruht auf Vorarbeiten, die vom Antragsteller und seinen MitarbeiterInnen im Rahmen des SFB 340, des BMBF-Verbundprojekts VERBMOBIL und des vom Ministerium für Wissenschaft und Kunst Baden-Württemberg geförderten Projekts “Entwicklung lexikalischer Wissensbasen” (ELWIS) durchgeführt worden sind.

Automatische inkrementelle Annotation

Im Rahmen des ELWIS-Projekts (Storrer, Feldweg und Hinrichs 1993, Feldweg und Hinrichs 1996) wurde in Zusammenarbeit mit dem IMS Stuttgart das bereits erwähnte Tag-Set STTS für das Deutsche entwickelt, das strikt auf syntaktischer Distribution beruht (Thielen und Schiller 1994, Schiller u.a. 1995). Durch die Arbeit mit verschiedenen Korpora und Annotationsverfahren wurde entsprechendes Know-How in der manuellen sowie der automatischen Formatkonversion erworben. Als Beispiel sei die Konversion der Annotationssysteme der Daten aus SADAW und CELEX in das STTS-Format angegeben. Der in (Feldweg 1995a) und (Feldweg 1995b) beschriebene stochastische Tagger *Likely* ist ein Standardwerkzeug der Wortartenannotation fürs Deutsche geworden.

Im Rahmen des Projekts B7 (Rooth/Abney) des SFB 340 (1995–1997) stand die Arbeit mit Korpora an zentraler Stelle, wobei Darstellungsformate für Chunks und für (partielle) syntaktische Strukturen definiert wurden. Die Arbeiten dieses Projekts betrachten neben dem Englischen auch das Deutsche, wobei vom STTS ausgegangen wird. Durch Arbeiten wie (Abney 1991), (Abney 1992), (Abney 1996a) und (Abney 1996b) ist die wichtige Funktion von Chunks für die linguistische Theoriebildung und die Architektur von Systemen zur Sprachverarbeitung begründet. In (Abney 1996b) wird ein Chunk-Parser vorgestellt, der in (Light 1996) in einer Anwendung aufs Deutsche dargestellt wird.
(Abney 1996a) und (Abney 1996b) zeigen, daß bei der Abbildung von einfachen auf komplexe linguistische Strukturen bereits sehr viel durch deterministische endliche Verfahren erreicht werden kann. In (Abney 1996a) findet sich auch eine umfassende Darstellung und Auswertung des Forschungsstands im Bereich der Annotationsmethoden auf Chunk- und Satzebene.

Im Tübinger Teilprojekt “Datensammlung und Korpusannotation” des BMBF Verbundprojekts Verbmobil-II werden Korpora gesprochener Sprache für das Deutsche, Englische und Japanische in sog. *Baumbanken* annotiert. Die Baumbank des Deutschen (Stegmann, Schulz und Hinrichs 1998) umfaßt gegenwärtig 16 000 vollständig annotierte Satzeinträge und basiert auf einem Annotationsschema, das auf der Wortklassenebene die Kategorien des STTS-Tagset und auf der Satzebene die Theorie der topologischen Felder verwendet. Diese auf syntaktischer Ebene größtenteils manuell erstellte Baumbank soll für die im Projektbereich “Automatische inkrementelle Annotation” zu entwickelnden Lernverfahren als Trainingsmaterial und für die Projektbereiche “Annotationserschließung” und “Theoriegeleitete Annotationsveränderung” als Datenmaterial eingesetzt werden.
Das Annotationssystem in (Stegmann u.a. 1998) hat zwei besondere Merkmale, die es grundlegend von anderen, zum Beispiel dem des NEGRA-Projekts, unterscheiden. Zum einen baut es direkt auf der Theorie der topologischen Felder auf. Die syntaktische Struktur enkodiert damit primär Stellungseigenschaften. Dependenzen sind hauptsächlich mittels Kno-

tenbezeichnungen ausgedrückt. Dadurch kommt die Annotation ohne kreuzende Äste bei diskontinuierlichen Konstruktionen aus. Zum anderen werden spezielle Label verwandt, um nichtauflösbare Ambiguitäten zu kennzeichnen. In der vergleichbaren Baumbank des NEGRA-Projekts muß in jedem Fall eine Entscheidung zugunsten einer Lesart getroffen werden (Skut, Krenn, Brants und Uszkoreit 1997).

Auch zur Anwendung von maschinellen Lernverfahren im Bereich der Syntax liegen bereits verschiedene Einzeluntersuchungen vor. In (Abney 1996c) und (Riezler 1996) wurden stochastische Phrasenstrukturgrammatiken um Merkmalsstrukturen erweitert. (Kübler 1996) vergleicht verschiedene Tagger hinsichtlich ihrer Anwendbarkeit aufs Deutsche, wobei der Schwerpunkt auf dem im Tagger von (Brill 1993b) enthaltenen Lernverfahren liegt. In (Kübler 1998) wird ein *memory-based*-Lernverfahren zur Induktion von syntaktischer Struktur für das Deutsche vorgestellt. Dieses Verfahren beinhaltet bereits implizit den im Projekt verfolgten inkrementellen Ansatz, da bereits gefundene Strukturen rekursiv die Grundlage für weitere Strukturannahmen bilden.

Annotationserschließung

Im Bereich Annotationserschließung wird es darum gehen, eine Baumbeschreibungssprache als Anfragesprache für Korpora zu entwickeln.
Eine wichtige eigene Vorarbeit in diesem Bereich stellt die bereits fertiggestellte Dissertation (Kallmeyer 1998a) dar, sowie einige Veröffentlichungen, die im Zusammenhang mit dieser Dissertation entstanden sind (Kallmeyer 1996, Kallmeyer 1997, Kallmeyer 1998b). Kallmeyer beschäftigt sich mit Baumbeschreibungssprachen, allerdings nicht unter der im Projekt A1 verfolgten Fragestellung. In ihrer Dissertation geht es darum, eine auf Baumbeschreibungen (im Gegensatz zu Strukturen) basierte Erweiterung von Tree Adjoining Grammars (TAG) zu definieren, die es erlaubt, Baumbeschreibungen mit unterspezifizierter Dominanz zu generieren und auf diese Weise geeignete unterspezifizierte Repräsentationen für Skopusambiguitäten zu erzeugen. Unter Dominanz wird hier die reflexive transitive Hülle der direkten Nachfolger-Beziehung verstanden, das heißt, ein Knoten u_1 dominiert einen Knoten u_2 gdw. entweder $u_1 = u_2$ gilt oder u_2 "unter" u_1 hängt. Neben der Möglichkeit der Unterspezifikation besitzt der entwickelte Grammatikformalismus Vorteile von TAG-ähnlichen Grammatiken, speziell den erweiterten Lokalitätsbereich und eine Beschränkung der generativen Kapazität im Sinne der sogenannten *schwach kontext-sensitiven* Grammatiken. Unter dieser Zielsetzung werden *Tree Description Grammars (TDG)* und, als weitere Variante, *lokale TDGs* definiert.
Der Definition von (lokalen) TDGs geht die Einführung einer Baumlogik voran, einer quantorenfreien Logik erster Ordnung, zusammen mit einem syntaktischen Folgerungsbegriff, für den Korrektheit und Vollständigkeit gezeigt werden. Die in Kallmeyers Arbeit als Elemente einer TDG zugelassenen Baumbeschreibungen, sogenannte *descriptions*, erlauben es, eine Unterspezifikation der Dominanz auszudrücken, was mit einer geeigneten Querysprache für syntaktisch annotierte Korpora auch auf jeden Fall möglich sein sollte. Die Ausdrucksmöglichkeiten von *descriptions* sind jedoch zu eingeschränkt für Anfragen an Korpora, da sie z.B. weder Negationen noch Disjunktionen erlauben. Die von Kallmeyer verwendete Baumlogik und der Begriff der *description* werden einer der Ausgangspunkte sein bei Überlegungen dazu, wie eine Baumbeschreibungssprache aussehen muß, die für Anfragen an syntaktisch

aufbereitete Korpora eingesetzt wird.

Theoriegeleitete Annotationsveränderung

Im Bereich der theorieorientierten Annotationsveränderung fließen Vorarbeiten aus verschiedenen Grundansätzen zusammen, die notwendig sind, um das sich neuartig und speziell im Kontext des SFB 1711 stellende Problem der Repräsentation detaillierter partieller Analysen sinnvoll angehen zu können. Die Problemstellung läßt sich in die zwei Teile *Repräsentation* und *Erschließung* gliedern.
Es ist eine notwendige Voraussetzung für die Definition einer partiellen Repräsentationform, daß eine strikte Unterscheidung zwischen Objekten und Beschreibungen getroffen wird. In dieser Hinsicht sind die Arbeiten von Paul King (King 1989, King 1994, King 1998) einschlägig. Auf diesen aufbauend wird in (Richter und Sailer 1997, Richter und Sailer in Vorb.) gezeigt, daß es unter bestimmten, wohl definierten Bedingungen möglich ist, üblicherweise als Beschreibungen interpretierte Ausdrücke als Objekte aufzufassen.
Im Bereich der Erschließung liegt zum einen praktische Erfahrung im Umgang mit verschiedenen Annotationssystemen für dieselbe Annotationsebene vor, die bei der Konvertierung von Daten aus dem SADAW- und CELEX-Format in das STTS im Rahmen des ELWIS-Projekts Eingang gesammelt wurde. Zum anderen haben sich Arbeiten wie (Hinrichs und Nakazawa 1996) und (Meurers 1995) intensiv mit dem Problem der *Matching Conditions* bei komplexen Abbildungssystemen wie Lexikalischen Regeln beschäftigt. Die Analogie zu Lexikalischen Regeln ist sehr treffend, da sich auch bei ihnen die Frage stellt, ob sie eine Abbildung zwischen Objekten oder Beschreibungen darstellen. In den genannten Arbeiten werden beide Ansichten exploriert.
Der empirische Bereich der Verbsyntax ist nicht nur für Partnerprojekte im SFB 1711 von großer Bedeutung (B3 und B10), sondern es liegt auch eine Reihe von Eigenarbeiten zum Deutschen (Hinrichs und Nakazawa 1989, Hinrichs und Nakazawa 1994, Hinrichs und Nakazawa 1997) und zu anderen Sprachen (Hinrichs, Kathol und Nakazawa 1997) vor. In (Meurers 1997) wurde eine Analyse traditionell schwer zu behandelnder, oft unbeachteter aber in Korpora attestierter Daten entwickelt. Gerade letztere Arbeit setzt einen Schwerpunkt auf den Vergleich zwischen traditionellen empirischen Arbeiten, theoretischen Ansätzen in GB und einer innovativen Umsetzung innerhalb einer HPSG-Fragmentgrammatik.

3.5 Ziele, Methoden, Arbeitsprogramm und Zeitplan

3.51 Ziele

Das übergeordnete Projektziel besteht darin, Textkorpora so zu annotieren und zu erschliessen, daß die Korpusdaten für sprachwissenschaftliche Einzeluntersuchungen fruchtbar gemacht werden können. Aus dieser theoriegeleiteten Annotation linguistischer Daten ergeben sich drei Schwerpunkte, die im Projekt gleichrangig und gleichzeitig bearbeitet werden sollen:

1. Automatische inkrementelle Annotation

   (a) Definition von wortübergreifenden Entitäten (Chunkebene)

(b) Erstellung einer Chunk-Grammatik für das Deutsche im Hinblick auf eine Anwendung des Chunk-Parsers von Abney (1996b)

(c) Testen verschiedener symbolischer Lernalgorithmen, um die Chunks gemäß der Theorie der topologischen Felder zu vollständigen Bäumen zu verbinden

(d) automatische Annotation eines speziellen Korpus

(e) Bereitstellung der Ergebnisse für andere Benutzer, insbesondere für andere Teilprojekte des SFB 1711

2. Annotationserschließung

   (a) Ermitteln der in linguistischen Repräsentationen (implizit oder explizit) enthaltenen strukturellen Information, die für linguistische Fragestellungen relevant ist

   (b) Auf Grundlage von (a) Entwicklung einer Baumbeschreibungssprache als intelligente Anfragesprache für syntaktisch annotierte Korpora

   (c) Untersuchung der formalen Eigenschaften der gewählten Baumbeschreibungssprache

   (d) Implementation eines Prototypen eines Query-Systems für ein spezielles Korpus

   (e) Bereitstellung der Ergebnisse für andere Anwender, insbesondere für andere Projekte des SFB

3. Theoriegeleitete Annotationsveränderung

   (a) exemplarische Spezifikation einer empirischen Domäne für die theoriegeleitete Datenerschließung.

   (b) Definition eines Abbildungsformalismus zwischen Repräsentationsformen/-arten.

   (c) Definition einer Deskriptionssprache über GB-basierte Strukturen mithilfe der in Bereich 2 entwickelten allgemeinen Baumbeschreibungssprache.

   (d) Instantiierung des Abbildungsformalismus zur Überführung von theorieneutralen Strukturen (Bereich 1) in GB-basierte Strukturskizzen.

3.52 Methoden und Arbeitsprogramm

Das Projekt gliedert sich in drei Bereiche, die in enger Abstimmung ein gemeinsames Großziel verfolgen, wobei jeder Bereich ein Modul bei der Gewinnung von Erkenntnissen aus und über Korpora darstellt. Das gemeinsame Ziel ist der Entwurf einer lückenlosen Kette von bloßem, unannotiertem, elektronisch verfügbarem Text bis hin zu einem für detaillierte, partielle linguistische Theorien aufgearbeiteten und intelligent abfragbaren Korpus. Die einzelnen beteiligten Module betrachten (i) die Annotation von bloßem Text zur syntaktischen Ebene, (ii) die Zugriffsmöglichkeiten auf annotierte Korpora und (iii) die Abbildbarkeit auf benutzerdefinierte, partielle Strukturen. Jedes dieser Module steht in methodologischem und inhaltlichem Zusammenhang mit anderen Partnerprojekten innerhalb des SFB 1711.

Automatische inkrementelle Annotation

Der erste Projektbereich befaßt sich mit der Fragestellung, wie Korpora automatisch annotiert werden können. Bei der Annotation müssen verschiedene linguistische Ebenen berücksichtigt werden. Auf der untersten Annotationsebene sollen Wortart-Informationen bereitgestellt werden; auf der nächsthöheren Ebene Wortgruppen, die durch syntaktische Beziehungen gekennzeichnet sind; und auf der obersten Ebene die syntaktische (Baum-)Struktur von Sätzen. Für diese Arbeiten werden die Verbmobil-Baumbank als Trainings- und Referenzmaterial sowie unannotierte Zeitungskorpora aus der ECI-Korpussammlung als zusätzliche Datenbasis dienen.
Die Strategie, die hier verfolgt werden soll, ist das maschinelle Lernen von syntaktischen Strukturzuweisungen. Mittels maschineller Lernverfahren lassen sich Inkonsistenzen ausschließen, für die eine manuelle Annotation anfällig ist. Vor allem beim POS-Tagging wurden bereits Ergebnisse erzielt, die mit anderen Verfahren vergleichbar sind (Brill 1993b).
Wie in 3.2 festgestellt, ist das Lernen der gesamten syntaktischen Annotation in einem Schritt ein unüberwindbares Problem. Die Komplexität der Aufgabe ergibt sich aus der rekursiven Natur der Konstituenten. Wird in einem der ersten Rekursionsschritte eine Teilkonstituente falsch angebunden, wird sich dieser Fehler auch auf die Anbindung der so entstandenen Konstituente negativ auswirken. Die Komplexität erhöht sich exponentiell relativ zur Länge des Satzes.
Um dieses Problem zu minimieren, soll hier ein inkrementeller Ansatz verfolgt werden. Der erste Schritt wird das Tagging mittels eines bereits existierenden Taggers sein. Die verwendeten Tagger (z.B. der von Brill (1993b)) werden dem SFB zur Verfügung gestellt. Nach dem Tagging wird eine Annotation auf der Chunk-Ebene durchgeführt. Erst im Anschluß werden Lernverfahren eingesetzt, um die Verknüpfung der Chunks miteinander zu einer Gesamtstruktur zu erreichen. Zum Zeitpunkt des Einsatzes von Lernverfahren ist ein Großteil der syntaktischen Analyse also bereits abgeschlossen, wodurch die Rekursivität der Aufgabe minimiert wird.

Das Aufgabengebiet in Bereich 1 läßt sich in zwei Bereiche unterteilen, nämlich die automatische Annotation der Daten auf der Chunk-Ebene und die Erweiterung der Chunk-Annotation zu einem vollständigen Baum, der den Kriterien der topologischen Felder genügt.

Annotation auf Chunk-Ebene

Auf wortübergreifender Ebene stehen sich prinzipiell zwei Konzepte gegenüber, das der (syntaktische) Konstituenten und das der Chunks. Für den Satz in (2a) ergeben sich dabei die folgenden Strukturen:

(2) a. (daß) die Beratungen über den Bebauungsplan am Montag aufgenommen werden

b. Konstituenten:
K1: *die Beratungen über den Bebauungsplan*
K2: *am Montag aufgenommen werden*
K11: *die*
K12: *Beratungen*
K13: *über den Bebauungsplan*
K21: *am Montag*
K22: *aufgenommen werden*
...

c. Chunks:
   C1: *die Beratungen*
   C2: *über den Bebauungsplan*
   C3: *am Montag*
   C4: *aufgenommen werden*

Die Konstituentenstruktur enthält sehr viel syntaktische und semantische Information (z.B. daß *über den Bebauungsplan* Teil der NP ist und daß *am Montag* das Verb modifiziert). Dies bedeutet jedoch, daß auch sehr viel Information vorhanden sein muß, um die korrekte Konstituentenstruktur eines Satzes erhalten zu können. In den Beiträgen in (Mehl, Mertens und Schulz 1996) wurden für den Bereich der PPs die Möglichkeiten und Grenzen bei der Erstellung von Konstituentenstrukturen dargestellt.
Chunkanalysen von Sätzen haben den Vorteil, daß Chunks nur diejenigen Wörter miteinander verbinden, für die aus offensichtlichen syntaktischen Kriterien geschlossen werden kann, daß sie Teil derselben Konstituente sind. Dies bedeutet, daß eine effiziente, verläßliche automatische Annotation auf Chunk-Ebene möglich ist. Es wird jedoch nichts über die rekursive Einbettung der Chunks untereinander gesagt, das heißt es ist nicht geklärt, in wieweit Chunks selbst wieder aus Chunks bestehen können oder sollten ([*der* [*über seine Wiederwahl*] *erfreute Vorsitzende*]).
Im Rahmen des Projektes soll die Definition von Chunks so erweitert werden, daß ihre empirische Verwendbarkeit gewährleistet ist. Es muß dabei zwischen linguistischen Anforderungen (z.B. Rekursivität) und der Komplexität der Anwendung abgewogen werden. Eine mögliche Vorgehensweise bei dem in den Sätzen (1) illustrierten Problem besteht darin, die Nominalphrase und die Adjektivphrase als separate Chunks zu behandeln, die in der folgenden Phase zusammengefügt werden.

Vervollständigung der Chunk-Annotation zu Bäumen

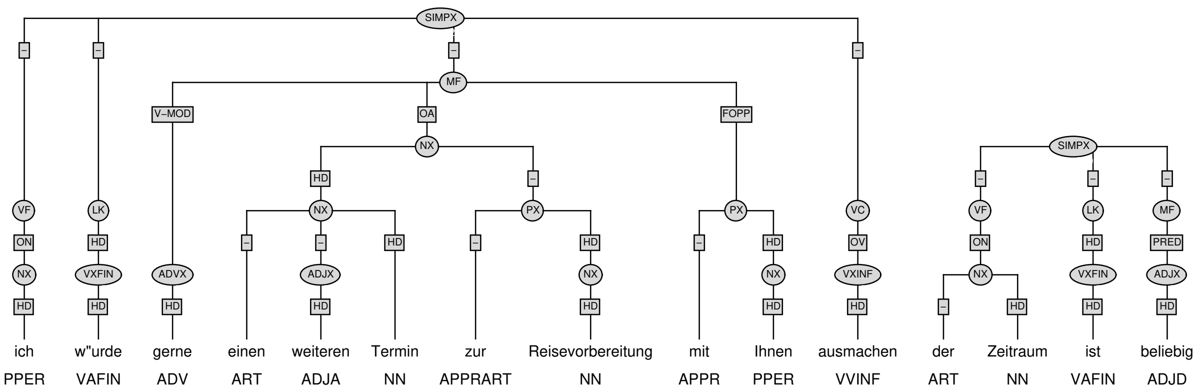
Bei einer näheren Betrachtung der Chunkanalyse zeigt sich, daß zum Teil Felder und Chunks zusammenfallen. So legt Beispiel (2c) nahe, daß der Verbalkomplex als ein Chunk aufgefaßt wird. Bei den anderen Beispielen lassen sich mehrere Chunks zu einem Feld zusammenfassen. Es bietet sich also an, die Chunkanalyse zu einer topologischen Felderanalyse zu erweitern.

(3) ich würde gerne einen weiteren Termin zur Reisevorbereitung mit Ihnen ausmachen [der Zeitraum ist beliebig]

In Satz (3) ergeben sich folgende Chunks:

(4) C1: *gerne*
    C2: *einen weiteren Termin*
    C3: *zur Reisevorbereitung*
    C4: *mit Ihnen*

Die Ermittlung des Vorfelds, der linken Klammer und des Verbkomplexes stellen kein Problem dar, da die einzelnen Wörter nur auf Felderebene projiziert werden müssen. Im Mittelfeld dagegen muß C3 an C2 angebunden werden, und anschließend müssen C1 und C4 mit der daraus entstehenden Struktur zum Mittelfeld projeziert werden.



Um diese Aufgabe automatisch zu erledigen, bieten sich maschinelle Lernverfahren an. Aufgaben, die von solchen Lernverfahren gelöst werden, können in zwei Hauptbereiche aufgeteilt werden: Segmentierung und Klassifizierung. Auch linguistische Probleme, wie z.B. die Vervollständigung der Chunk-Parses zu vollständigen Bäumen lassen sich als solche Probleme beschreiben. Die Segmentierung besteht hier aus der Eingrenzung der im nächsten Schritt zusammenzufassenden Chunks. Die Klassifizierung besteht in der Ermittlung der korrekten Bezeichnung für diese neue Struktur. Oft ist es sinnvoll, die Aufgaben der Segmentierung und der Klassifizierung zur selben Zeit zu lösen. Ist z.B. bekannt, daß C2 zu einer Nominalphrase projiziert wird, liegt damit auch die Information vor, daß eine Präpositionalphrase wie C3 angebunden werden kann. Ist diese Information bei der Segmentierung nicht vorhanden, gestaltet sich diese Entscheidung wesentlich schwieriger. Diejenigen unter 3.2 genannten Lernverfahren, die die Integration linguistischen Wissens erlauben, ermöglichen eine Verbindung der beiden Teilaufgaben.
Da durch die Verbmobilbaumbank zwar bereits ein großes Trainingskorpus vorhanden ist, aber keine explizite Grammatik, liegen die notwendigen Ressourcen für Entscheidungsbaum-Modelle, *memory-based learning* und *transformation-based error-driven learning* vor. Als Ausgangspunkt für die Projektarbeit soll zunächst der *memory-based*-Ansatz herangezogen werden, da hier gute Ergebnisse bei nichtlokalen Beziehungen vorliegen und Zugriff auf lexikalische Eigenschaften möglich ist, was beispielsweise im Bereich der PP-Anbindung notwendig ist.

Annotationserschließung

Im Zentrum des zweiten Projektbereichs steht die Aufgabe, eine intelligente Anfragesprache für syntaktisch annotierte Korpora zu entwickeln. Hierbei ist zunächst zu klären, welche in syntaktischen Annotationen enthaltene Information linguistisch relevant ist, mit anderen Worten, welche Anfragen an ein syntaktisch annotiertes Korpus idealerweise möglich sein sollten. Anschließend soll basierend auf einer geeigneten Baumlogik eine Anfragesprache entwickelt werden. Die gewählte Baumbeschreibungssprache hängt nicht nur von den Anforderungen an eine solche Sprache ab, vielmehr bestimmen auch technische und formale Gründe wünschenswerte Eigenschaften der Anfragesprache, die die Art der zuzulassenden Anfragen eingeschränken.
In gewisser Weise kann man eine Anfragesprache als eine Art Filter betrachten, der (wie auch die Annotation selbst) den Zugriff auf verschiedene Aspekte der Daten einschränkt. Unter

der Voraussetzung einer geeigneten Annotation können durch eine Anfragesprache einzelne Eigenschaften der Daten explizit gemacht werden, während andere Eigenschaften nicht sichtbar sind, da sie in der Anfragesprache nicht ausgedrückt werden können. Insofern sollte eine Anfragesprache auf einer abstrakteren Ebene zusammen mit der Annotation die (angestrebte) theoretische Sicht des Linguisten auf die von ihm betrachteten Daten widerspiegeln. Im Idealfall führen formale Erkenntnisse darüber, was in einer Baumbeschreibungssprache ausgedrückt werden kann, dazu, daß zuvor nicht berücksichtigte Datenaspekte erschlossen werden.

Betrachtet man Anfragesprachen für Korpora wie das in 3.2 beschriebene Stuttgarter System Xkwic oder das im Saarbrücker SFB 378 entwickelte Querysystem von Duchier, so wird deutlich, daß diese Anfragesprachen weitgehend durch praktische bzw. anwendungsorientierte und weniger durch mathematische Gesichtspunkte motiviert sind. In beiden Fällen handelt es sich nicht um Baumlogiken im eigentlichen Sinne, da, wie in 3.2 schon erwähnt, für diese Sprachen keine denotationelle Semantik definiert ist und daher keine direkte modelltheoretische Verbindung zwischen einem Ausdruck in der Anfragesprache und den gesuchten syntaktischen Strukturen besteht. Hinzu kommt, daß die erlaubten Anfragen hinsichtlich ihrer Syntax und Semantik nicht immer homogen sind. Xkwic erlaubt beispielsweise Kombinationen von regulären Ausdrücken mit verschiedenen Anweisungen, die teilweise die zu suchenden Strukturen weiter einschränken und teilweise Anweisungen an das Suchprogramm sind.

Im Gegensatz dazu soll im Rahmen des hier beschriebenen Projektbereiches als Anfragesprache eine Baumlogik bzw. Baumbeschreibungssprache entwickelt werden, die mit einer denotationellen modelltheoretischen Semantik versehen ist. Das heißt, der Bezug zwischen Baumbeschreibungen und Strukturen ist über eine Erfüllbarkeitsrelation definiert, wobei endliche Bäume die intendierten Modelle sind. Voraussetzung dafür, daß eine Struktur eine Beschreibung erfüllt, ist daher, daß diese Struktur die spezifischen strukturellen Eigenschaften von endlichen Bäumen aufweist (zur Axiomatisierung von endlichen Bäumen siehe z.B. (Backofen, Rogers und Vijay-Shanker 1995), (Kallmeyer 1998a)). Das bedeutet, daß man aufgrund der Baumlogik aus einer gegebenen Anfrage auf weitere Eigenschaften der gesuchten Strukturen schließen kann. Dies ist mit den beiden oben genannten Anfragesystemen nicht möglich.

Die Baumbeschreibungssprache soll unabhängig sein von Überlegungen, die den Suchvorgang selber betreffen. Das heißt, diese Sprache stellt gegenüber bestehenden Anfragesprachen eine weitere Abstraktionsstufe dar. Die Baumbeschreibungssprache soll nicht nur als Anfragesprache verwendet werden, sondern sie wird auch im Projektbereich 3 dazu dienen, Relationen zwischen Annotationen zu formulieren.

Die Ausdrücke einer solchen Baumbeschreibungssprache sind *Constraints*, die über endlichen Bäumen interpretiert werden. Mit anderen Worten, ein solcher Constraint beschreibt eine Menge von Bäumen. Wenn für eine bestimmte Annotation festgestellt werden soll, ob diese Annotation eine gegebene Anfrage erfüllt, so entspricht dies der Frage, ob die Annotation in der von der Anfrage beschriebenen Baumsprache ist, das heißt für die jeweilige Anfrage (als constraint-basierte Grammatik betrachtet) muß das Wortproblem gelöst werden. Daher sind, unabhängig von einer konkreten Implementation, formale Eigenschaften der Baumbeschreibungssprache von großem Interesse, speziell Resultate zur Entscheidbarkeit und Komplexität. Diese Fragen sollen im Rahmen des Projektes untersucht werden.

Bei der Frage, wie ausdrucksstark die Baumbeschreibungssprache sein soll bzw. für welche

linguistische Information ein Zugriff mit Hilfe der Anfragesprache möglich sein muß, werden insbesondere die spezifischen Überlegungen derjenigen Projekte im SFB mit einbezogen, die das in dem hier beschriebenen Projektbereich entwickelte Query-System verwenden wollen. Dies sind zum einen Projekte, die auf syntaktisch annotierte Korpora zugreifen wollen, nämlich die Projekte B1, B3, B4, B8 und B10. Zum anderen ist geplant, die Anfragesprache auch im Projekt B5 zur Erschließung prosodischer Annotationen einzusetzen. Hier handelt es sich zwar nicht um syntaktische Bäume, aber es geht um Annotationen, die struktureller Natur und daher auch durch eine Baumbeschreibungssprache zu erschließen sind. Die prosodische Annotation der in B5 verwendeten Korpora ist eine Baumstruktur, die in der Regel durch eine entsprechende Klammerung repräsentiert wird.
Als exemplarisches, syntaktisch annotiertes Korpus wird bei der Überlegung dazu, welche strukturellen Eigenschaften linguistisch interessant sind, die auch im Projektbereich 1 verwendete Verbmobil-Baumbank betrachtet. Daneben werden sich die gewünschten Eigenschaften der Anfragesprache an den Bedürfnissen von Projektbereich 3 orientieren. Da es sich hier um Strukturüberführungen handelt, sind andere linguistische Merkmale relevant als für das Ziel eines linguistisch adäquaten Anfragesystems.
Es ist offenkundig von großem Interesse, nicht nur auf der Ebene der Ketten von Wörtern, sondern auch bezogen auf die syntaktische Struktur, Information aus einem Korpus zu extrahieren. Eine für linguistische Datenerschließung geeignete Anfragesprache muß es daher erlauben, sich in einer Anfrage auf strukturelle Gegebenheiten zu beziehen. Das heißt, daß es möglich sein muß, nach bestimmten Teilstrukturen zu suchen.
Wie in Abschnitt 3.2 schon erläutert, ist ein Query-System, das wie Xkwic auf regulären Ausdrücken basiert, auf Anfragen beschränkt, in denen die lineare Abfolge der Teilausdrücke festgelegt ist. Das folgende Beispiel zeigt jedoch, daß es wünschenswert ist, in einer Anfrage nicht nur Teilstrukturen vorgeben zu können, was in einem gewissen Sinne die gesuchten Strukturen nur partiell beschreibt und damit unterspezifiziert, sondern darüber hinaus sogar einzelne Eigenschaften dieser Teilstrukturen offen bzw. unterspezifiziert zu lassen. Insbesondere sollte auch eine Unterspezifikation der linearen Abfolge zulässig sein.
In (5) wird eine Anfrage danach gestellt, in welcher Skopusreihenfolge intersektive und nicht-intersektive Adjektive als Modifikatoren innerhalb einer NP auftreten. Die Anfrage ist durch ein Constraint einer an (Rogers und Vijay-Shanker 1994) und (Kallmeyer 1998a) angelehnten Logik ausgedrückt. Dabei denotiert $\lhd$ die Relation *parent* oder *immediate dominance*, also die direkte Nachfolgerbeziehung zwischen Knoten in einem Baum, und $\lhd^*$ denotiert Dominanz (*dominance*). Die Konstante $\prec$ denotiert lineare Präzedenz zwischen Knoten. Die freien Variablen x_i, $i \in \mathbb{N}$, denotieren Knoten. Diese Knoten können mit Attribut-Wert Paaren versehen sein, wobei $a(x_i) \approx v$ bedeutet, daß der von x_i denotierte Knoten mit dem Attribut-Wert Paar $\langle a, v\rangle$ versehen ist.

$$(5) \quad \begin{gathered} x_1 \lhd^* x_2 \wedge x_2 \lhd x_3 \wedge x_2 \lhd x_4 \wedge x_3 \prec x_4 \wedge x_1 \lhd^* x_5 \\ \wedge\, x_5 \lhd x_6 \wedge x_5 \lhd x_7 \wedge x_6 \prec x_7 \wedge x_4 \lhd^* x_8 \wedge x_7 \lhd^* x_8 \\ \wedge\, cat(x_1) \approx N \wedge cat(x_2) \approx N \wedge cat(x_3) \approx AP \wedge intersec(x_3) \approx yes \\ \wedge\, cat(x_4) \approx N \wedge cat(x_5) \approx N \wedge cat(x_6) \approx AP \wedge intersec(x_6) \approx no \\ \wedge\, cat(x_7) \approx N \wedge cat(x_8) \approx N \end{gathered}$$

Die lineare Präzedenz der beiden Adjektive ist in (5) unterspezifiziert, da weder $x_3 \prec x_6$ noch $x_6 \prec x_3$ ausgeschlossen ist. Anfragen dieser Art kann man in einem System wie Xkwic nur stellen, indem man eine Disjunktion angibt, also im Prinzip für jede der möglichen linearen

Abfolgen eine eigene Anfrage stellt. Eine echte Unterspezifikation der linearen Abfolge ist nicht möglich.
In dem oben betrachteten Beispiel ergibt sich die Unterspezifikation der linearen Abfolge aus der Unterspezifikation der Dominanzbeziehung zwischen vollständig festgelegten Teilstrukturen. Derartige Baumbeschreibungen, die eine Unterspezifikation der Dominanzbeziehung zwischen Knoten zulassen, finden sich in Arbeiten zur unterspezifizierten Semantik, speziell zu Skopusambiguitäten, wie z.B. die in (Muskens 1995) vorgeschlagene *underspecified logical form* und die in (Kallmeyer 1997) verwendeten *descriptions*. Greift man die Ideen dieser Arbeiten auf, so bietet es sich an, als Baumbeschreibungssprache eine einfache Baumlogik erster Ordnung zu verwenden, wie etwa in (Rogers 1994) und (Rogers und Vijay-Shanker 1994) im Zusammenhang mit Tree Adjoining Grammars (TAG) vorgeschlagen oder wie in (Kallmeyer 1996, Kallmeyer 1998b) für Tree Description Grammars (TDG) verwendet. Derartige Baumsprachen werden im Projekt als Ausgangspunkt für Überlegungen zur Gestalt einer Query-Sprache für syntaktisch annotierte Korpora dienen. Die hier genannten quantorenfreien Baumlogiken erster Ordnung werden für die Entwicklung einer Anfragesprache erweitert werden müssen. Beispielsweise werden, wie es auch bei Duchier der Fall ist, Existenz- und Allquantifikation hinzugefügt werden, um Eindeutigkeit ausdrücken zu können.
Eine Logik erster Ordnung mit Quantoren wird jedoch nicht genügen. Die im Projektbereich 3 angestrebten Strukturüberführungen sollen eine Abbildung relativ flacher Strukturen wie in der Verbmobil-Baumbank in komplexere GB-nahe Strukturen ermöglichen. Dementsprechend muß die Baumbeschreibungssprache auch die Formulierung von GB-spezifischen strukturellen Zusammenhängen ermöglichen, z.B. die Suche nach Ketten. (6) zeigt die von Rogers (1994) vorgeschlagene Definition eines Prädikates *Chain* in $L^2_{K,P}$, der schon in 3.2 erwähnten monadischen Prädikatenlogik zweiter Ordnung. $x_i, i \in \mathbb{N}$ sind hier wie in (5) Variablen erster Ordnung, und X ist eine Variable zweiter Ordnung, denotiert also eine Menge (bzw. ein einstelliges Prädikat).

(6)
$$\begin{aligned}
Chain(X) \quad \equiv \quad & (\exists! x_1)[X(x_1) \wedge Target(x_1)] \wedge (\exists! x_2)[X(x_2) \wedge Base(x_2)] \\
& \wedge (\forall x_3)[X(x_3) \wedge \neg Target(x_3) \\
& \qquad \rightarrow (\exists! x_4)[X(x_4) \wedge Link(x_4, x_3)]] \\
& \wedge (\forall x_5)[X(x_5) \wedge \neg base(x_5) \\
& \qquad \rightarrow (\exists! x_6)[X(x_6) \wedge Link(x_5, x_6)]] \\
& \wedge (\forall x_7, x_8)[X(x_7) \wedge (Link(x_7, x_8) \vee Link(x_8, x_7)) \\
& \qquad \rightarrow X(x_8)]
\end{aligned}$$

Die Bedeutung von (6) ist, daß eine Kette eine Menge von Knoten ist, die genau einen Zielknoten (*target*) und genau einen Ausgangsknoten (*base*) enthält, bei der alle Knoten außer dem Ziel einen eindeutigen Antezedensknoten und alle außer dem Ausgangsknoten einen eindeutigen Vorgänger besitzen (die Verbindung von einem Knoten zu seinem Nachfolger in der Kette wird durch die Relation *Link* ausgedrückt), und die maximal ist bezüglich der Relation *Link*.
In einem GB-nah annotierten Korpus muß eine Suche nach Ketten möglich sein. Dieses Beispiel zeigt daher, daß eine Sprache erster Ordnung wie in dem Querysytem von Duchier verwendet zwar einen ersten Schritt in Hinblick auf intelligentere Anfragesprachen darstellt, daß es jedoch erforderlich ist, über die Ausdrucksstärke der von Duchier benutzten Querysprache hinauszugehen. Neben der monadischen Prädikatenlogik zweiter Ordnung sollen

auch andere Logiken höherer Ordnung in Betracht gezogen werden, beispielsweise die in 3.2 erwähnten *Context constraints over finite trees* aus (Niehren u.a. 1997a, Niehren u.a. 1997b). Speziell bei der Verwendung von Logiken höherer Ordnung muß man abwägen zwischen der aus linguistischen Gründen gewünschten Ausdrucksstärke auf der einen Seite und der Komplexität der gewählten Logik auf der anderen Seite. Aus formaler Sicht ist eine Mindestvoraussetzung, daß die gewählte Logik entscheidbar ist, das heißt, daß gewährleistet ist, daß sich für eine gegebene Anfrage und eine Annotation immer in endlich vielen Schritten entscheiden läßt, ob die Annotation die Anfrage erfüllt oder nicht. Ein zweiter wichtiger, formaler Gesichtspunkt betrifft die Komplexität. Eventuell wird man sich bei der angestrebten Ausdrucksstärke einschränken müssen, um eine zu hohe Komplexität zu vermeiden oder um Entscheidbarkeit zu gewährleisten. Abgesehen von der Wahl der Logik werden formale Eigenschaften wie Entscheidbarkeit auch durch die Wahl der intendierten Modelle beeinflußt. Auch wenn im Bereich 2 zunächst von endlichen Bäumen ausgegangen wird, könnte man sich auch andere Strukturen als Annotationen vorstellen, wie beispielsweise baumähnliche Annotationen, die ein Überkreuzen von Kanten zulassen, wie sie im Saarbrücker Projekt NEGRA verwendet werden. An dieser Stelle besteht eine enge Beziehung zum Projekt A2, das sich mit dem Zusammenhang zwischen der Wahl eines Datentypen und den sich daraus ergebenden formalen Eigenschaften beschäftigt.

Ist eine adäquate Baumbeschreibungssprache entwickelt, so soll unter Verwendung dieser Sprache zunächst exemplarisch für die Verbmobil-Baumbank der Prototyp eines Query-Systems implementiert werden. Dieses System wird den anderen Projekten des SFB 1711 zur Verfügung gestellt. In einem weiteren Schritt wird die entwickelte Anfragesprache mit dem Ergebnis aus Projektbereich 1 kombiniert, so daß sich ein System ergibt, das es erlaubt, ein gegebenes Korpus zunächst automatisch syntaktisch zu annotieren und anschließend Anfragen an dieses Korpus zu stellen.

Theoriegeleitete Annotationsveränderung

Während in Bereich “Automatische inkrementelle Annotation” die automatische Annotation großer Datenmengen in einem theorieneutralen Repräsentationsformat angestrebt wird und Bereich “Annotationserschließung” die Grundlagen für die direkte Extraktion linguistischer Information in Form einer Querysprache schafft, beschäftigt sich der dritte Projektbereich mit der prinzipiellen Möglichkeit der Abbildbarkeit eines gegebenen Korpus in ein neues Korpus. Hierbei ist das neue Korpus jedoch qualitativ anders als bereits existierende Korpora, es enthält theorieorientierte partielle Analysen. Im Falle eines solchen Korpus soll von einem *Phänokorpus* gesprochen werden, da linguistische Theorien in der Regel an einem bestimmten Phänomen orientiert sind.

Ein syntaktisch annotiertes Korpus ist Grundvoraussetzung für eine gezielte Suche nach syntaktischen Konstruktionen. Um solche Suchen tatsächlich durchführen zu können, muß eine Querysprache bereitgestellt werden, die es ermöglicht, auf die in der Annotation gegebene Information Bezug zu nehmen. Für linguistische Fragestellungen ist ein solches Korpus allerdings sowohl zu informationsarm als auch zu informationsreich. Es enthält zu wenig Information für diejenigen Aspekte, die direkt mit der konkreten, zu untersuchenden Fragestellung zu tun haben. Es enthält jedoch zugleich auch zu viel Information, da es von Linguisten nicht erwartet werden kann, für jeden Aspekt aller im Korpus auftretenden Daten eine vollständi-

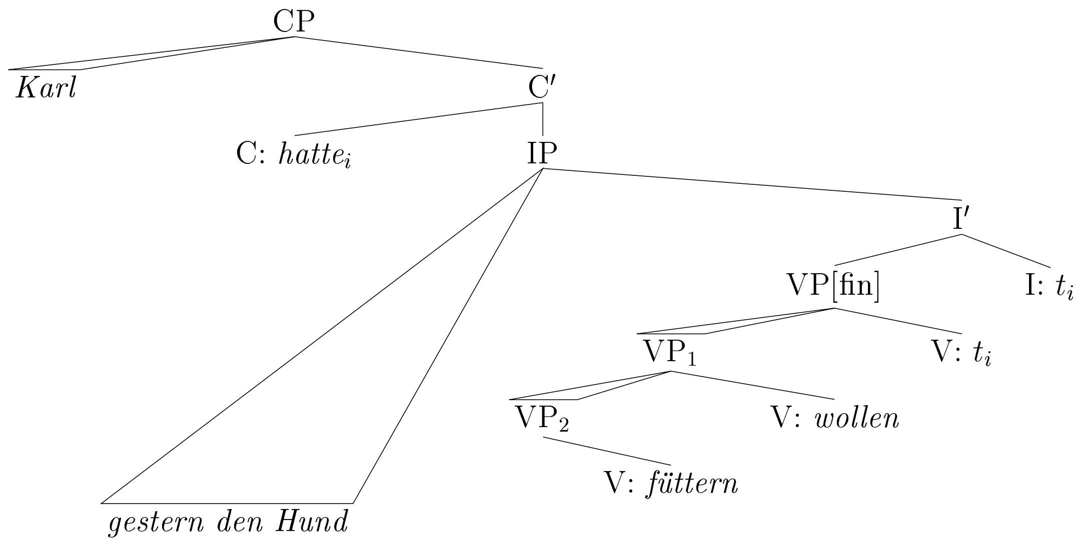
ge Analyse zu haben. Der Gegensatz ergibt sich aus dem prinzipiellen Kontrast zwischen vollständiger Analyse von Teilaspekten von Daten in der Linguistik und robuster Grobanalyse, die für die Erfassung großer Datenmengen unumgänglich ist. Neben Information, die nur in einem der beiden Systeme benötigt wird, gibt es auch Information, die zwar in beiden vorhanden ist, aber auf unterschiedliche Weise repräsentiert wird.
Aus der skizzierten Problemstellung ergibt sich, daß ein Korpus durch eine Anfrage gefiltert und für die gefilterten Daten eine an den Bedürfnissen einer bestimmten linguistischen Theorie ausgerichtete Beschreibung ausgegeben wird. Aus der Partialität theoriegeleiteter Darstellungen folgt notwendig, daß die Ausgabe eines Feinannotationsprozesses kein echter Korpus sein kann, d.h., kein Korpus im Sinne einer Interpretation eines Annotationssystems. Es liegen also statt vollständiger Objekte lediglich unterspezifizierte Beschreibungen solcher Objekte vor. Eine solche Menge von Beschreibungen bezeichnen wir als *Phänokorpus*, die ihn konstituierenden Baumbeschreibungen als *Baumskizzen*. Die Tatsache, daß Baumskizzen in der Linguistik sinnvoll sind, ist zum Beispiel dadurch dokumentiert, daß in jeder theoretischen Arbeit in den graphischen Darstellungen nichtrelevante Aspekte der Struktur ausgeblendet werden.
Im dritten Projektbereich wird die allgemeine Problemstellung am Phänomenbereich Verbsyntax des Deutschen exemplifiziert und konkretisiert. Der Phänomenbereich umschließt die Stellung des finiten Verbs sowie die Struktur des Verbkomplexes. Die Projekte B3 (Ehrich/Reis) und B10 (von Stechow) werden in dieser empirischen Domäne jeweils aus ihrer spezifischen Sicht theoretische Ansätze evaluieren und entwickeln.
An einem konkreten Beispiel aus dem Bereich Verbkomplex soll die Relevanz der allgemeinen Fragestellung verdeutlicht werden. Als Ausgangsannotationssystem und -korpus verwendet das Projekt die Verbmobil-Baumbank. Für die Betrachtung der Verbsyntax ist die Frage nach dem Rektionsstatus eines Verbs relevant. Im Verbmobil-Annotationssystem wird jedoch nicht zwischen "normalen" Infinitiven und sogenannten Ersatzinfinitiven unterschieden. Um diese Information hinzuzufügen, muß auf die Dependenzstruktur innerhalb des Verbkomplexes sowie auf Information aus zusätzlichen Quellen wie beispielsweise Selektionsrahmen für Hilfsverben zugegriffen werden.
Als Beispiel von in beiden Systemen vorhandener Information kann die Position des finiten Verbs herangezogen werden. So enthält das Verbmobil-Annotationssystem Information über vorangestellte finite Verben. Sie sind der Kopf des Satzes und werden direkt von der Kategorie „Linke Klammer" dominiert. In einer eher GB-orientierten Analyse hätte diese Information die Gestalt, daß das finite Verb in C^0 steht und mit einer Spur in I^0 und einer Spur im Kopf der höchsten V-Projektion koindiziert ist.
In (7) ist ein Beispiel für einen Satz angegeben, der Verbbewegung und einen kleinen Verbkomplex enthält. In (8) ist die Strukturskizze angegeben, in die der Satz unter der hier angenommenen Feinannotation abgebildet werden soll. (9) ist die Strukturskizze als Baumgraphik dargestellt. Die in (9) auftretenden Dreiecke bezeichnen die Dominanzrelation.

(7) Karl hatte gestern den Hund füttern wollen.

(8)
- Die allgemeinen Prinzipien auf wohlgeformte Strukturen: Appropriateness bzw. Angabe der Einbettungsreihenfolge der funktionalen Projektionen: CP $\lhd^*$ IP $\lhd^*$ VP[finit] $\lhd^*$ VP[$\neg$ finit]).
- Die Angabe des Inhalts der funktionalen Köpfe (C^0: *hatte*$_i$; I^0: t_i, V^0: t_i).

- Die Dominanz der übrigen Verben: VP[fin] $\lhd^*$ V: t_i; VP[fin] $\lhd^*$ VP_1;VP_1 $\lhd$ V: *wollen*; VP_1 $\lhd^*$ VP_2; VP_2 $\lhd$ V: *füttern*.

(9) Strukturskizze:



Dieses konkrete Beispiel ist als Instanz eines allgemeinen Abbildungsformalismus zu verstehen. Mathematisch sei hierbei eine Abbildungsfunktion F angenommen, die parametrisiert ist bezüglich einer Menge V von Abbildungsvorschriften. Eine Abbildungsvorschrift in V hat die Form $\delta \Longrightarrow \delta'$, wobei δ eine Beschreibung der Beschreibungssprache des Inputkorpus ist, und δ' eine Beschreibung der Beschreibungssprache des Zielannotationssystems. Die Funktion F_V ist dann wie folgt definiert:

(10) Sei I eine Interpretation eines Annotationssystems A_1, $\mathcal{D}_1$ die Menge der A_1-Beschreibungen, sei $\mathcal{D}_2$ die Menge der Beschreibungen eines Annotationssystems A_2, sei V eine Menge von Paaren $\delta \Longrightarrow \delta' \in \mathcal{D}_1 \times \mathcal{D}_2$, dann ist für jedes $u \in I$,

$$F_V(u) := \left\{ \bigwedge_i \delta'_i \;\middle|\; \begin{array}{l} \text{es gibt ein } \delta_i \Longrightarrow \delta'_i \in V, \\ \text{so daß } u \in [\![\; \delta_i \;]\!]_1 \end{array} \right\}$$

Aus der Definition von F_V folgt, daß durch modulare Erweiterung der Vorschriftenmenge V die resultierende Strukturskizze entsprechend angereichert werden kann.
Strukturskizzen als Beschreibungen werden mit Hilfe der Baumbeschreibungssprache aus Projektbereich 2 definiert.
Für das obige Beispiel (7) ergibt sich die Einbettung der funktionalen Projektionen aus einer allgemeinen Abbildungsvorschrift, deren linke Seite von jedem Satz im Eingabekorpus erfüllt wird. Die Verbspuren werden durch eine Regel erzeugt, deren linke Seite ein Verb in der linken Satzklammer beschreibt. Bei der Beschränkung auf Verbsyntax wird keine Regel auftreten, die etwas über die interne Struktur nominaler Projektionen aussagt. Diese erscheinen in der Ausgabeskizze als maximal unterspezifiziert hinsichtlich ihrer syntaktischen Anbindung. In (9) ist das große Dreieck so zu lesen, daß die einzige vorhandene Information die ist, daß jedes Wort/jeder Knoten im Baum, der unter der Basis eines Dreiecks auftrit, dominiert

wird von dem syntaktischen Knoten an der Spitze des Dreiecks. Aus dieser Definition ergibt sich beispielsweise auch, daß die NP *den Hund* von VP_2 dominiert sein könnte.
Phänokorpora stellen einerseits die einzige Möglichkeit dar, automatisch von robusten Annotationssystemen zu analysespezifischen Daten zu gelangen. Andererseits ist nicht von vornherein klar, wie beispielsweise ein Abfragesystem auf einem Phänokorpus arbeiten kann. Ein solches Querysystem muß auf einer Metasprache über Beschreibungen basieren, wie sie beispielsweise in Arbeiten zu Generalisierungen im Lexikon in HPSG angenommen wird. Alternativ dazu könnte man versuchen, Strukturskizzen den Status von Objekten zuzuerkennen. Als Beispiel mag die graphische Darstellung der Strukturskizze zu (7) dienen. Hierzu wird die Signatur für Bäume dahingehend erweitert, daß Dreiecke Bäume sind, und daß es auf Objektebene zwei Arten von ,,Elternbeziehung" gibt: die Standardbeziehung der unmittelbare Dominanz zwischen zwei Knoten, und eine, die der unterspezifizierten Dominanz entspricht. Zu einer solchen Baumsignatur kann nun wiederum eine entsprechende Querysprache definiert werden.
Durch den Übergang von Objekten zu Beschreibungen wird deutlich, daß die in Bereich 3 behandelten Verfahren hauptsächlich bei theorieorientierter Annotationsveränderung Anwendung finden. Bei inkrementeller Annotation, wie sie in Bereich 1 zentral ist, kann und soll auf Objektebene geblieben werden. Der Schritt von Struktur*skizzen* zu tatsächlichen — aber relativ wenig Information tragenden — Stukturen ermöglicht die systematische Erschließung der bei theorieorientierter Annotationsveränderung gewonnenen Daten mit den in Bereich 2 untersuchten Mitteln.
Die drei Projektbereiche decken damit einerseits das Bedürfnis von Korpusbenutzern ab, sich eine Repräsentationsebene und ein Repräsentationsformat selbst zu definieren, andererseits wird durch die Untersuchung der Methoden der Korpusverarbeitung (inkrementelle Annotation, Query und theoriegeleitete Annotation) gezeigt, wie die Kluft zwischen theoretischen und empirischen Daten überbrückt werden kann.

3.53 Zeitplan

1999

- Automatische inkrementelle Annotation
    * Erstellung der Chunkgrammatik
    * Prüfung, Akquisition und Einarbeitung in projektrelevante Software
- Annotationserschließung
    * Bedarfsprüfung für die zu definierende Querysprache
    * Erste Definition
- Theoriegeleitete Annotationsveränderung
    * Grammatiktheoretische Untersuchungen zum Bereich Verbkomplex
    * Definition einer Abbildungsfunktion

2000

- Automatische inkrementelle Annotation
    * Absicherung der Annotationsqualität für Tag- und Chunkebene

  - ∗ Anwendung eines Lernverfahrens zur inkrementellen Erstellung einer Baumbank
- – Annotationserschließung
  - ∗ Definition einer Querysprache
  - ∗ Untersuchung der formalen Eigenschaften der Querysprache
- – Theoriegeleitete Annotationsveränderung
  - ∗ Formalisierung des partiellen Repräsentationsschemas
  - ∗ Einpassen der Querysprache in den Abbildungsformalismus

2001

- – Automatische inkrementelle Annotation
  - ∗ Optimierung des Lernverfahrens
  - ∗ Abschluß offener Arbeiten und Dokumentation
- – Annotationserschließung
  - ∗ Implementierung des Prototyps eines Querysystems
  - ∗ Abschluß offener Arbeiten und Dokumentation
- – Theoriegeleitete Annotationsveränderung
  - ∗ Definition von Zugriffsmöglichkeiten innerhalb eines Phänokorpus
  - ∗ Abschluß offener Arbeiten und Dokumentation

3.6 Stellung innerhalb des Programms des Sonderforschungsbereichs

Das Verhältnis von Empirie und Theorie in der Grammatikforschung als zentraler Forschungsgegenstand des beantragten SFBs soll im Teilprojekt A1 am Gegenstand von elektronisch verfügbaren Textkorpora untersucht werden. Die Ergebnisse dieser Untersuchungen sollen den Teilprojekten des SFBs, bei denen die datenorientierte Untersuchung linguistischer Einzelphänomene im Vordergrund steht, mit dem Ziel zur Verfügung gestellt werden, die Validität der vom Projekt A1 entwickelten Verfahren und Repräsentationssysteme zur Korpusannotation empirisch zu überprüfen.

Verbindung zu anderen Teilprojekten
Hinsichtlich der in A1 zu untersuchenden Repräsentationssysteme zur Korpusannotation und der zu entwickelnden Verfahren der Datenerschließung ergeben sich vielfältige Querverbindungen zu anderen Teilprojekten:

A2 (Mönnich): Die Projekte A2 und A1 beschäftigen sich beide mit Repräsentationssystemen für linguistische Daten. Aus den im Projekt A2 geplanten Untersuchungen zu abstrakten Datenstrukturen ergeben sich enge Verbindungen zu den modelltheoretischen Grundlagen der in A1 untersuchten Baumbeschreibungssprache. Durch die von A2 geplanten Untersuchungen zum Warao sind universalgrammatische Implikationen für linguistische Annotationssysteme zu erwarten, die in A1 berücksichtigt werden sollen.

A3 (Sternefeld): Die Projekte A3 und A1 beschäftigen sich beide mit theorieneutralen und theoriegeleiteten Annotationen von Korpusdaten. Die von A3 experimentell erhobenen Daten stellen eine wichtige Ergänzung zu den von A1 annotierten Primärdaten dar. Die von A3 untersuchten suboptimalen syntaktischen Strukturen des Deutschen ermöglichen eine Robustheitsprüfung des in A1 verwendeten Annotationsschemas hinsichtlich negativer Daten.

B1 (Berger), B4 (Eschbach-Szabó) und B8 (Raecke): In diesen Projekten sollen annotierte Korpora erstellt und verwendet werden. Hierbei werden sich sprachübergreifende Gesichtspunkte der Korpusannotation hinsichtlich der verwendeten linguistischen Kategorien und der von A1 entwickelten Verfahren der Korpusannotation und -erschließung ergeben.

B3 (Ehrich/Reis), B5 (Féry/Hamm) und B10 (von Stechow): Die drei Projekte beschäftigen sich ebenso wie A1 schwerpunktmäßig mit deutschen Daten. Dadurch sind sie potentielle Interessenten für die in A1 entwickelten Verfahren zur Erschließung deutscher Korpora. Andererseits sind durch phänomenzentrierten Untersuchungen in diesen drei Projekten wichtige Impulse für die von A1 verwendeten Repräsentationssysteme und für die Entwicklung der Querysprache (einschließlich möglicher Generalisierung auf prosodische Strukturen [B5]) zu erwarten. Eine zusätzliche Gemeinsamkeit mit dem Projekt B3 besteht hinsichtlich des Phänomenbereichs Verbkomplex.

Z (Zentralprojekt): Die vom Zentralprojekt akquirierten Korpora und Softwaretools sollen auf ihre Verwendbarkeit für die Belange von A1 überprüft und soweit wie möglich angewendet werden. Der vom Projekt Z organisierte inhaltliche Austausch zwischen Teilprojekten hinsichtlich Korpusarbeiten soll durch die Bereitstellung der in A1 entwickelten Verfahren nachhaltig unterstützt werden.

Verbindung zu Forschergruppen außerhalb Tübingens

Neben der Zusammenarbeit mit anderen Teilprojekten des SFB 1711 sollen Kooperationen mit in- und ausländischen Forschergruppen etabliert bzw. fortgesetzt werden, die ebenfalls auf dem Gebiet der Korpuslinguistik arbeiten:

- Center for Dutch Language and Speech, Antwerpen: Im Rahmen des EG-Projekts "Learning Computational Grammars" im TMR-Programm der Europäischen Kommission bestehen Kontakte zu Prof. Dr. W. Daelemans (Antwerpen/Tilburg), die vor allem hinsichtlich des Bereichs Automatische inkrementelle Annotation ausgebaut werden sollen.

- SFB 378, Teilprojekt C3 (NEGRA): Das im NEGRA-Projekt entwickelte Annotationswerkzeug wird für die in A1 betrachteten Baumbanken verwendet. Gleichzeitig soll der Gedankenaustausch über die komplementären Forschungsziele hinsichtlich performanzorientierter (C3, SFB 378) und kompetenzgrammatischer (A1, SFB 1711) Auswertung von Korpusdaten fortgesetzt werden.

- AT&T Laboratories, New Jersey: Die Kooperation mit Herrn Dr. Steven Abney, einem ehemaligen Mitarbeiter am Lehrstuhl Allgemeine Sprachwissenschaft/Computerlinguistik, in den Bereichen des Chunkparsings und des Einsatzes stochastischer Methoden in der Computerlinguistik soll fortgesetzt werden.

Literatur

Abney, S.: 1991, Parsing by chunks, *in* R. Berwick, S. Abney und C. Tenney (Hrsg.), *Principle-Based Parsing*, Kluwer Academic Publishers.

Abney, S.: 1992, Prosodic structure, performance structure and phrase structure, *Proccedings, Speech and Natural Language Workshop*, Morgan Kaufmann Publisher, San Mateo, CA, S. 425 – 428.

Abney, S.: 1996a, Part-of-speech tagging and partial parsing, *in* K. Church, S. Young und G. Bloothooft (Hrsg.), *Corpus-Based Methods in Language and Speech*, Kluwer Academic Publishers.

Abney, S.: 1996b, Partial parsing via finite-state cascades, *in* J. Carroll (Hrsg.), *Workshop on Robust Parsing (ESSLLI '96)*, S. 8 – 15.

Abney, S.: 1996c, Stochastic attribute-value grammars, The Computation And Language E-Print Archive cmp-lg/9610003.

Argamon, S., Dagan, I. und Krymolowski, Y.: 1998, A memory-based approach to learning shallow natural language patterns, *Proceedings of the 17th COLING.*

Backofen, R., Rogers, J. und Vijay-Shanker, K.: 1995, A first-order axiomatization of the theory of finite trees, *Journal of Logic, Language and Information* **4**(1), S. 5 – 39.

Bech, G.: 1955, Studien über das deutsche verbum infinitum, Det Kongeliege Danske Videnskaber Selskab; Dan. Hist. Filol. Medd. Bind 35, no. 2, 1955; Bind 36, no. 6, 1957. reprint: Niemeyer, Tübingen, 1983.

Blackburn, P. und Meyer-Viol, W.: 1994, Linguistics, logic and finite trees, **2**, S. 2 – 39.

Bod, R. und Scha, R.: 1996, Data-Oriented Language Processing. An Overview, *Technical Report LP-96-13*, Institute for Logic, Language and Computation, Universiteit van Amsterdam.

Bos, J.: 1995, Predicate logic unplugged, *Proceedings of the 10th Amsterdam Colloquium.*

Brants, T. und Skut, W.: 1998, Automation of treebank annotation, *Proceedings of NeMLaP3/CoNLL98. New Methods in Language Processing and Computational Natural Language Learning, Australia*, S. 49 – 57.

Brill, E.: 1993a, Automatic grammar induction and parsing free text: a transformation-based approach, *Proceedings of the* ARPA *Workshop on Human Language Technology*, S. 237 – 242.

Brill, E.: 1993b, *A Corpus-Based Approach to Transformation-Based Learning*, Doktorarbeit, University of Pennsylvania.

Brill, E. und Resnik, P.: 1994, A rule-based approach to prepositional phrase attachment disambiguation, *Proceedings of COLING.*

Carpenter, B.: 1992, *The Logic of Typed Feature Structures*, Cambridge University Press. Cambridge, Massachusetts, USA.

Carpenter, B. und Penn, G.: 1996, Efficient parsing of compiled typed attribute value logic grammars, *in* H. Bunt und M. Tomita (Hrsg.), *Recent Advances in Parsing Technology*, Kluwer.

Charniak, E.: 1996, Tree-bank grammars, *Technical Report CS-96-02*, Department of Computer Science, Brown University.

Christ, O.: 1994, A modular and flexible architecture for an integrated corpus query system, *Proceedings of COMPLEX'94*.

Christ, O. und Schulze, B. M.: 1996, Ein flexibles und modulares Anfragesystem für Textkorpora, *in* H. Feldweg und E. W. Hinrichs (Hrsg.), *Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Niemeyer, Tübingen, S. 121 – 133.

Church, K. W. und Mercer, R. L.: 1993, Introduction to the special issue on computational linguistics using large corpora, *Computational Linguistics* **19**(1), S. 1 – 24.

Cutting, D., Kupiec, J., Pederson, J. und Sibun, P.: 1992, A practical part-of-speech tagger, *Third Conference on Applied Natural Language Processing*, ACL, S. 133 – 140.

Daelemans, W.: 1995, Memory-based lexical acquisition and processing, *in* P. Steffens (Hrsg.), *Machine Translation and the Lexicon*, Vol. 898 of *Lecture Notes in Artificial Intelligence*, Springer, S. 85 – 98.

Daelemans, W.: 1998, Abstraction is harmful in language learning, *Proceedings of NeMLaP3/CoNLL98. New Methods in Language Processing and Computational Natural Language Learning, Australia*, p. 1.

den Besten, H. und Edmondson, J. A.: 1983, The verbal complex in continental West Germanic, *in* W. Abraham (Hrsg.), *On the Formal Syntax of the Westgermania*, number 3 in *Linguistik aktuell*, Amsterdam/Philadelphia, S. 155–216.

EAGLES: 1996a, Detailed overview of existing annotation practices. EAGLES document EAG-TCWG-SASG/3.

EAGLES: 1996b, Morphosyntactic annotation. EAGLES document EAG-CSG/IR-T3.1.

EAGLES: 1996c, Syntactic annotation: Survey of annotation practices. EAGLES document EAG-TCWG-SASG/2.

Eisenberg, P.: 1989, *Grundriß der deutschen Grammatik*, Stuttgart. 2., überarbeitete und erweiterte Auflage.

Engel, U.: 1970, Regeln zur Wortstellung, *in* U. Engel (Hrsg.), *Forschungsberichte des Instituts für deutsche Sprache*, Vol. 5, o.O., S. 7 – 148.

Evers, A.: 1975, *The Transformational Cycle in Dutch and German*, PhD thesis, Rijksuniversiteit Utrecht.

Feldweg, H.: 1995a, Implementation and evaluation of a German HMM for POS disambiguation, *EACL Sigdat Workshop*, Dublin.

Feldweg, H.: 1995b, Stochastische Wortartendisambiguierung des Deutschen, *in* H. Feldweg und E. Hinrichs (Hrsg.), *Lexikon & Text*, Niemeyer, Tübingen, S. 241 – 254.

Feldweg, H. und Hinrichs, E.: 1996, *Lexikon & Text. Wiederverwertbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Niemeyer, Tübingen.

Haegeman, L. und Riemsdijk, H. v.: 1986, Verb projection raising, scope and the typology of verb movement rules, *Linguistic Inquiry* **17**, S. 417 – 466.

Haider, H.: 1993, *Deutsche Grammatik — generativ. Vorstudien zu einer projektiven Grammatik*, Gunter Narr, Tübingen.

Hinrichs, E. und Nakazawa, T.: 1989, Flipped out: Aux in German, *Papers from the 25th Regional Meeting of the Chicago Linguistic Society*, Vol. 25, CLS, Chicago, Illinois, S. 193 – 202.

Hinrichs, E. und Nakazawa, T.: 1994, Partial-VP and split-NP topicalization in German — an HPSG analysis, *in* E. Hinrichs, D. Meurers und T. Nakazawa (Hrsg.), *Partial-VP and Split-NP Topicalization in German — An HPSG Analysis and its Implementation*, Arbeitspapiere des SFB 340, Nr. 58, Universität Tübingen, S. 1 – 46.

Hinrichs, E. und Nakazawa, T.: 1996, Applying lexical rules under subsumption, *Proceedings of COLING 96*, COLING 96, Copenhagen, S. 543 – 549.

Hinrichs, E. und Nakazawa, T.: 1997, Third construction and VP extraposition in German — an HPSG analysis, in (Hinrichs, Kathol und Nakazawa 1997).

Hinrichs, E., Kathol, A. und Nakazawa, T. (Hrsg.): 1997, *Complex Predicates in Nonderivational Syntax*, Syntax and Semantics Series, Academic Press, San Diego.

Hinrichs, E., Meurers, D., Richter, F., Sailer, M. und Winhart, H.: 1997, *Ein HPSG-Fragment des Deutschen, Teil 1: Theorie*, Arbeitspapiere des SFB 340, Nr. 95, Universität Tübingen.

Höhle, T. N.: 1986, Der Begriff "Mittelfeld". Anmerkungen über die Theorie der topologischen Felder, *in* A. Schöne (Hrsg.), *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen 1985*, Vol. 3, S. 329 – 340.

Kallmeyer, L.: 1996, Tree Description Grammars, *in* D. Gibbon (Hrsg.), *Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference*, Mouton de Gruyter, Berlin.

Kallmeyer, L.: 1997, A syntax-semantics interface with Synchronous Tree Description Grammars, *Proceedings of Formal Grammar*, Aix-en-Provence.

Kallmeyer, L.: 1998a, *Tree Description Grammars and Underspecified Representations*, Doktorarbeit, Universität Tübingen. erscheint in Arbeitspapiere des SFB 340.

Kallmeyer, L.: 1998b, Underspecification in Tree Description Grammars, *in* H. P. Kolb und U. Mönnich (Hrsg.), *The Mathematics of Syntactic Structures*, Mouton de Gruyter. im Erscheinen.

Kathol, A.: 1995, *Linearization-Based German Syntax*, Doktorarbeit, Ohio State University.

King, P. J.: 1989, *A logical Formalism for Head-Driven Phrase Structure Grammar*, Doktorarbeit, University of Manchester.

King, P. J.: 1994, An expanded logical formalism for Head-Driven Phrase Structure Grammar, *Arbeitspapiere des SFB 340 59*, Universität Tübingen.

King, P. J.: 1998, Towards truth in HPSG, unveröffentlichtes Manuskript, Universität Tübingen.

Kiss, T.: 1993, *Infinite Komplementation. Neue Studien zum deutschen Verbum infinitum*, Theorie des Lexikons. Arbeiten des Sonderforschungsbereichs 282 Nr. 42, Bergische Universität Gesamthochschule Wuppertal.

Kroch, A. S. und Santorini, B.: 1991, The derived constituent structure of the West Germanic verb-raising construction, *in* R. Freidin (Hrsg.), *Principles and Parameters in Comparative Grammar*, MIT-Press, S. 269 – 338.

Kübler, S.: 1996, *Automatischer Erwerb von Wortklassifizierungsregeln*, Magisterarbeit, Universität Trier.

Kübler, S.: 1998, Learning a lexicalized grammar for German, *Proceedings of NeMLaP3/CoNLL98. New Methods in Language Processing and Computational Natural Language Learning, Australia*, S. 11 – 18.

Light, M.: 1996, CHUNP: Partial parsing and underspecified representations, *Proceedings of the ECAI-96 Workshop: Corpus-Oriented Semantic Analysis.*

Magerman, D.: 1995, Statistical decision-tree models for parsing, *Proceedings of the 33rd Annual Meeting of the ACL*, S. 276 – 283.

Mehl, S., Mertens, A. und Schulz, M.: 1996, Präpositionalsemantik und PP-Anbindung, *Schriftenreihe Informatik Bericht Nr. SI-16*, Gerhard-Mercator-Universität Duisburg.

Meurers, W. D.: 1995, Towards a semantics for lexical rules as used in HPSG, *Proceedings of the Conference on Formal Grammar*, Barcelona. Auch in: Proceedings of the ACQUILEX II Workshop on Lexical Rules, 1995, Cambridge, UK.

Meurers, W. D.: 1997, Statusrektion und Wortstellung in kohärenten Infinitkonstruktionen des Deutschen, in (Hinrichs, Meurers, Richter, Sailer und Winhart 1997).

Morawietz, F. und Cornell, T.: 1997, Representing constraints with automata, *Proceedings of ACL*, Madrid.

Muskens, R.: 1995, Order-independence and underspecification, *in* J. Groenendijk (Hrsg.), *Ellipsis, Underspecification, Events and More in Dynamic Semantics*, DYANA Report R2.2.C.

Neumann, G.: 1997, Applying explanation-based learning to control and speeding-up natural language generation, *Proceedings of the ACL-EACL 1997*, S. 214 – 221.

Niehren, J., Pinkal, M. und Ruhrberg, P.: 1997a, On Equality Up-to Constraints over Finite Trees, Context Unification, and One-Step Rewriting, *Proceedings of the International Conference on Automated Deduction*, Lecture Notes in Computer Science.

Niehren, J., Pinkal, M. und Ruhrberg, P.: 1997b, A Uniform Approach to Underspecification and Parallelism, *Proceedings of ACL*, Madrid.

Pollard, C. und Sag, I.: 1987, *Information Based Syntax and Semantics. Vol.1: Fundamentals*, CSLI Lecture Notes 13.

Ramshaw, L. und Marcus, M.: 1995, Text chunking using transformation-based learning, *Proceedings of the Third Workshop on Very Large Corpora*, S. 82 – 94.

Reyle, U.: 1993, Dealing with ambiguities by underspecification: Construction, representation and deduction, *Journal of Semantics* **10**(2),S. 123 – 179.

Richter, F. und Sailer, M.: 1997, Underspecified semantics in HPSG, *in* H. Bunt, L. Kievit, R. Muskens und M. Verlinden (Hrsg.), *Proceedings of the Second International Workshop on Computational Semantics, 8–10 January 1997, Tilburg*, S. 234 – 246.

Richter, F. und Sailer, M.: in Vorbereitung., Lf constraints on terms of ty2, *in* A. Przepiórkowski und R. Borsley (Hrsg.), *Slavic in HPSG*, CSLI.

Riezler, S.: 1996, Quantitative extensions of constraint logic grammars, Universität Tübingen.

Rogers, J.: 1994, *Studies in the Logic of Trees with Applications to Grammar Formalisms*, Doktorarbeit, University of Delaware.

Rogers, J.: 1996, On descriptive complexity, language complexity and GB, *in* P. Blackburn und M. de Rijke (Hrsg.), *Specifying Syntactic Structures*, CSLI Publications.

Rogers, J. und Vijay-Shanker, K.: 1994, Obtaining trees from their descriptions: an application to Tree-Adjoining Grammars, *Computational Intelligence* **10**(4), S. 401 – 421.

Samuelsson, C.: 1994, *Fast Natural Language Parsing Using Explanation-Based Learning*, Doktorarbeit, Swedish Institute of Computer Science, Kista.

Schiller, A., Teufel, S. und Thielen, C.: 1995, Guidelines für das Tagging deutscher Textkorpora mit STTS, Manuskript Universität Stuttgart und Universität Tübingen.

Skut, W., Krenn, B., Brants, T. und Uszkoreit, H.: 1997, An annotation scheme for free word order languages, *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP), Washington, D.C.*

Stechow, A. v. und Sternefeld, W.: 1988, *Bausteine syntaktischen Wissens*, Westdeutscher Verlag.

Stegmann, R., Schulz, H. und Hinrichs, E.: 1998, Stylebook for the German Treebank in VERBMOBIL, Draft, Universität Tübingen.

Storrer, A., Feldweg, H. und Hinrichs, E.: 1993, Korpusunterstützte Entwicklung lexikalischer Wissensbasen, *Sprache und Datenverarbeitung* **17**(2), S. 59 – 72.

Thielen, C. und Schiller, A.: 1994, Ein kleines und erweitertes Tagset fürs Deutsche, *in* H. Feldweg und E. Hinrichs (Hrsg.), *Lexikon & Text*, Niemeyer, Tübingen, S. 215 – 226.