3.1 Allgemeine Angaben zum Teilprojekt A1

3.1.1 Thema:

Repräsentation und Erschließung linguistischer Daten

3.1.2 Fachgebiete und Arbeitsrichtung:

Computerlinguistik, Syntax des Deutschen, Korpuslinguistik, Annotationsverfahren, Evaluation

3.1.3 Leiter:

Hinrichs, Erhard, Prof. Dr.
geb. am 17.8.1954
Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstr. 113
72074 Tübingen
Tel.: 07071 - 297 5446
Fax: 07071 - 551 335
E-mail: eh@sfs.nphil.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 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 (Ergänzungsausstattung)

Haushalts-jahr	Personalkosten	Sächl. Verw.-ausgaben	Investitionen	gesamt
bis 1998	0	0	0	0
1999	169.2	10.3	0	179.5
2000	186.0	8.5	0	194.5
2001	186.0	8.5	0	194.5
Zwischen-summe	541.2	27.3	0	568.5
2002	177.6	20.7	0	198.3
2003	177.6	4.5	0	182.1
2004	177.6	4.5	0	182.1

(Beträge in TDM)

3.2 Zusammenfassung

Das Verhältnis von Empirie und Theorie in der Grammatikforschung als zentraler Forschungsgegenstand des Sonderforschungsbereichs wird im Teilprojekt A1 am Gegenstand von elektronisch verfügbaren Textkorpora untersucht.

In der ersten Projektphase (1999-2001) wurden in einer hybriden Parsingarchitektur, die *finite-state*-Methoden mit maschinellen Lernverfahren des *memory-based learning* verbindet, robuste computerlinguistische Methoden entwickelt, mit deren Hilfe Textkorpora morpho-syntaktisch und syntaktisch automatisch annotiert werden können. Die so annotierten Korpora stellen eine neue Datenklasse für linguistische Einzeluntersuchungen dar und stehen als empirische Grundlage der linguistischen Theoriebildung zur Verfügung, die von anderen Teilprojekten des SFB 441 genutzt werden kann. Im Sinne der Wiederverwendbarkeit der annotierten Daten wurde darauf geachtet, dass das verwendete Annotationsschema theorieneutralen Charakter hat und sich in die gewünschte Zielrepräsentation einer spezifischen linguistischen Theorie jeweils übersetzen läßt. Um die annotierten Daten leichter zugänglich zu machen, ist gleichzeitig ein Korpusabfragesystem VIQTORYA mit einer logisch wohl-fundierten Abfragesprache entstanden, dessen graphische Benutzeroberfläche es ermöglicht, Suchanfragen menügesteuert zu formulieren und annotierte Korpora benutzerfreundlich zu recherchieren.

Aufbauend auf den o.g. Ergebnissen soll in der hier beantragten Projektphase die Bandbreite der verwendeten computerlinguistischen Annotationsverfahren und der untersuchten grammatischen Phänomenbereiche signifikant erweitert werden sowie die

Güte der Annotationen durch quantitative und phänomenorientierte Evaluationsverfahren systematisch überprüft werden.

Hinsichtlich der Gesamtthematik des SFB 441 zum Verhältnis von Empirie und Theorie in der Grammatikforschung möchte das Projekt einen Beitrag zu folgenden übergeordneten Fragestellungen leisten:

1. Welche Arten von linguistischer Information sind nötig, um zentrale grammatische Phänomene, wie etwa die Erkennung topologischer Felder oder grammatischer Funktionen im Deutschen, theorieneutral zu erschließen?

2. Welche computerlinguistischen Analyseverfahren eignen sich für die korpuslinguistische Erschließung grammatischer Phänomene?

3. Welche Evaluations- und Suchverfahren eignen sich, um die phänomenorientierte Annotation linguistischer Daten systematisch zu überprüfen?

3.3 Stand der Forschung

Für eine Darstellung der bisherigen Ergebnisse des Projekts verweisen wir auf den Arbeits- und Ergebnisbericht. Hier fassen wir nur diejenigen Ergebnisse kurz zusammen, die für die Projektarbeit in der hier beantragten Förderphase zielführend sind:

1. In der ersten Projektphase (1999-2001) wurde eine hybride Parsingarchitektur entwickelt, die *finite-state*-Methoden mit maschinellen Lernverfahren des *memory-based learning* verbindet. Dieser hybride Ansatz, der syntaktische Annotation als einen inkrementellen Prozess versteht, hat sich in ersten quantitativen Evaluationen als außerordentlich tragfähig erwiesen und soll in der kommenden Projektphase nach methodischen Gesichtspunkten, die den Stand der Forschung in der datenorientierten Computerlinguistik reflektieren, weiter ausdifferenziert werden. Der Stand der Forschung für die vorgesehenen Parsing- bzw. Lernverfahren wird in Abschnitt 3.3.1 näher skizziert.

2. Mit dem Verbalkomplex im Deutschen ist in der ersten Projektphase ein linguistischer Phänomenbereich des Deutschen exemplarisch hinsichtlich einer theorieneutralen syntaktischen Repräsentation und deren Überführung in die Spezifikationen von linguistischen Einzeltheorien untersucht worden. In der kommenden Projektphase sollen weitere für die Makrostruktur deutscher Sätze zentrale grammatische Phänomene hinzukommen, deren Relevanz für die (computer-)linguistische Forschung in Abschnitt 3.3.2 beschrieben wird.

3. Mit VIQTORYA steht ein Query-Tool zur Verfügung, dessen prototypische Implementierung einer logisch fundierten Abfragesprache es schon jetzt ermöglicht, nach komplexen syntaktischen Strukturen in großen, syntaktisch annotierten Textkorpora präzise zu suchen. In der kommenden Projektphase soll

die Implementierung der VIQTORYA zugrundeliegenden Query-Sprache in enger Kooperation mit dem Teilprojekt C1 und orientiert an den Bedürfnissen der phänomenorientierten Einzelprojekte erweitert werden. Das Query-Tool stellt gleichzeitig eine wichtige Voraussetzung für eine qualitative, phänomenorientierte Evaluation von automatischen Annotationsverfahren dar. Der Stand der Forschung auf dem Gebiet von Abfragesprachen und Query-Tools sowie von quantitativen und qualitativen Evaluationsverfahren in der Computerlinguistik wird in Abschnitt 3.3.3 näher beschrieben.

3.3.1 Lern- und Parsingverfahren in der datenorientierten Computerlinguistik

Der Stand der Forschung in der datenorientierten Computerlinguistik ist durch eine große methodische Vielfalt gekennzeichnet. Neben "klassischen" regelbasierten Ansätzen kommen in zunehmendem Maße Verfahren des maschinellen Lernens und probabilistische Parsingverfahren zur Anwendung.

***Finite-state*-Formalismus** Das im Teilprojekt A1 verfolgte inkrementelle Verfahren, einen auf *finite-state*-Methoden basierten Chunkparser (Abney, 1996) als ersten syntaktischen Analyseschritt zur Erkennung nicht-rekursiver Strukturen der Erkennung komplexerer syntaktischer Einheiten vorzuschalten, hat sich mittlerweile in der Forschung insgesamt bewährt und durchgesetzt.

Eine derartige Chunkanalyse, die die Anbindung von Phrasen zu komplexeren syntaktischen Einheiten zunächst offen läßt, kann als computerlinguistisches Pendant zum Ansatz unterspezifizierter Repräsentationen in Syntax und Semantik betrachtet werden, der in der theoretischen Sprachwissenschaft zunehmend an Bedeutung gewonnen hat (Bos, 1995; Pinkal, 1996; Reyle, 1993; Richter und Sailer, 1999).

Für das Englische haben Finch und Mikheev (1997) einen Chunker auf der Grundlage von *finite-state*-Methoden entwickelt und für das Französische Aït-Mokhtar und Chanod (1997). Für das Deutsche beinhalten das Verbmobil-System (Hinrichs et al., 2000) und das System ResoPPat (Clematide und Volk, 2001) eine Voranalyse durch einen Chunker. Das Informationserschließungssystem SMES (Neumann und Piskorski, 2001) und das Valenzextraktions-System von Wauschkuhn (1999) enthalten sowohl einen Chunker als auch einen Teilsatzgrenzenerkenner als Voranalyse. Das im Projekt A1 entwickelte Chunking-System ist jedoch das einzige, das eine nicht auf Textsorten eingeschränkte, robuste, an linguistischen Kriterien orientierte Annotation als Grundlage für eine volle linguistische Annotation bietet.

Memory-based learning Das in der ersten Projektphase eingesetzte Lernverfahren beruht auf dem Ansatz des *memory-based learning* (MBL). Hierbei handelt es sich um ein überwachtes Lernverfahren, bei dem die Trainingsinstanzen ohne Abstraktion in einer Instanzenbasis gespeichert werden. Die Klassifizierung neuer Instanzen erfolgt über die Suche nach den ähnlichsten Instanzen in der Instanzenbasis, deren Klassifizierung dann übernommen wird.

Da es sich bei *memory-based learning* um ein reines Klassifikationsverfahren handelt, wurde es auf dem Gebiet des Syntaxlernens bisher nur zum Lernen von Teilaufgaben eingesetzt. Veenstra (1998) und Cardie und Pierce (1998) lernen NP Chunks, Buchholz et al. (1999) verwenden ein kaskadiertes MBL-System, um grammatische Relationen zu annotieren. Argamon et al. (1998) bzw. Krymolowski und Dagan (2000) verwenden im Memory-Based Sequence Learning Fenster von variabler Länge, um nicht-rekursive Phrasen zu annotieren.

Im Gegensatz zu den obigen partiellen Verfahren wurde im Projekt der Ansatz verfolgt, vollständiges Parsing als Klassifikationsaufgabe zu definieren. Bedingt durch diese Herangehensweise musste der zugrundeliegende MBL-Algorithmus so adaptiert werden, dass die lineare Abfolge der als Merkmale dienenden Wörter und PoS-Tags ebenso berücksichtigt werden konnte wie auch die interne Struktur der Klassen, d.h. der kompletten Bäume. Um auch eine Generalisierung auf neue Eingaben zu erreichen, wurde außerdem eine Erweiterung eingeführt, die es erlaubt, die gefundenen ähnlichsten Bäume an die Eingabe anzupassen. Die Resultate wurden in Kübler und Hinrichs (2000,2001a,b) publiziert.

Konnektionistische Lernverfahren Ein weiteres Lernverfahren, das in der Kognitionswissenschaft und Künstlichen Intelligenz bereits auf eine Vielzahl von Fragestellung angewendet worden ist, stellen konnektionistische Ansätze dar. Es handelt sich hierbei ebenfalls um Klassifikationsverfahren, die in überwachte und unüberwachte Methoden eingeteilt werden können. Im Bereich der Sprachverarbeitung werden im Allgemeinen jedoch überwachte Verfahren vorgezogen, da so besser gewährleistet werden kann, dass speziell die intendierten Regularitäten gelernt werden.

Konnektionistische Systeme bestehen aus einer großen Anzahl einfacher Einheiten (Zellen, Neuronen), die sich Informationen in Form von Aktivierung der Zellen über gerichtete Verbindungen zusenden (Rumelhart und McClelland, 1986; Zell, 1994). Bei überwachten Lernverfahren wird dabei mit Hilfe unterschiedlicher Lernalgorithmen iterativ versucht, eine Konfiguration der Verbindungsgewichte zu finden, um das Zuordnungsproblem zwischen vorgegebenen Ein- und Ausgabemustern möglichst optimal zu lösen.

Was die Wahl der Architekturen und Lernalgorithmen anbelangt, besteht bei konnektionistischen Verfahren (im Gegensatz zum MBL) ein großer Freiheitsgrad. So stehen verschiedene Arten von vorwärtsgerichteten oder rekurrenten Netzwerkarchitekturen zur Verfügung, die jeweils mit unterschiedlichen Lernalgorithmen trainiert werden können. Bei diesen Lernalgorithmen handelt es sich im Allgemeinen um lokale Optimierungsmethoden, wie z.B. Backpropagation und Varianten davon. Für einen systematischen Vergleich verschiedener Backpropagation-Varianten für die Aufgabe der Graphem-Phonem-Zuordnung (*NETtalk*) siehe Steiner (1995).

Konnektionistische Verfahren wurden bisher für computerlinguistische Aufgabenstellungen, wie z.B. Grammatikalitätsentscheidungen (Lawrence et al., 2000), Graphem-Phonem-Zuordnungen (Stoianov und Nerbonne, 1999) oder das Lernen von semantischen Rollen (Miikkulainen und Mayberry, 1999) eingesetzt. Ein spezieller konnektionistischer Ansatz wird von Roth und Zelenko (1998) für das Part-of-Speech Tagging und von Muñoz et al. (1999) für die Annotation von nicht-rekursiven Nominalphrasen und Subjekt-Verb-Mustern eingesetzt. Der hierbei verwendete Lernalgorithmus basiert auf multiplikativer Gewichtsänderung.

Probabilistische kontextfreie Grammatiken Probabilistische kontextfreie Grammatiken (*probabilistic context free grammars*, PCFGs) bieten eine gute Ergänzung zu konnektionistischen Verfahren und zum *memory-based learning*, da sie hinsichtlich der Abstraktion von den Trainingsdaten einen Mittelweg einschlagen und beim Training zwar abstrahieren, jedoch Teilstrukturen der Daten in Form von kontextfreien Grammatikregeln übernehmen.

Für die automatische syntaktische Annotation wurden bereits vielfach Methoden eingesetzt, die auf PCFGs basieren. Bei den am weitesten entwickelten erweiterten PCFGs, die für die tiefe syntaktische Analyse eingesetzt werden, ist der Untersuchungsgegenstand jedoch meist die englische Sprache (Collins, 1999; Eisner, 2000; Charniak, 1997). Für das Deutsche steht bisher die Anwendung von PCFGs für die partielle Analyse syntaktischer Strukturen im Vordergrund (Skut und Brants, 1998; Brants, 1999; Schmid und Schulte im Walde, 2000), da die Modellierung der freieren Wortstellung des Deutschen neue Probleme aufwirft (Plaehn, 1999). Für die Bewältigung des Problems der knappen Daten, das beim überwachten Lernen und einer begrenzten Menge von Trainingsdaten eine große Rolle spielt, hat sich das *smoothing* mit Hilfe des *maximum-entropy*-Ansatzes bewährt (Charniak, 1999; Ratnaparkhi, 1999). Außerdem gibt es viel versprechende Hinweise, dass die begrenzte Menge annotierter Trainingsdaten mittels unüberwachter Lernverfahren und einer großen Menge nicht oder teilannotierter Trainingsdaten ergänzt werden kann und sich dieses Verfahren auf die syntaktische Annotation mit Hilfe von PCFGs erfolgreich anwenden lässt (Pereira und Schabes, 1992; Schmid und Schulte im Walde, 2000; Beil et al., 1999).

Bisher ist allerdings kein Ansatz bekannt, der sich die Regularitäten innerhalb der freieren Wortstellung des Deutschen zu Nutze macht, und auch weitere Wissensquellen wie Valenzlexika mit einbezieht, um basierend auf erweiterten PCFGs eine tiefe syntaktische Analyse mit der für das Englische erreichten Güte durchzuführen.

Daten- und Wissensquellen Alle Verfahren, die in der beantragten Projektphase zur Anwendung kommen, sind auf umfangreiche Daten- und Wissensquellen angewiesen. Dem Projekt stehen drei syntaktisch (teil-)annotierte Korpora zur Verfügung. Das NEGRA-Korpus umfasst 176.000 Token manuell annotierter Zeitungstexte (Frankfurter Rundschau). Aufgrund der unterschiedlichen Annotations-Schemata wird für die Nutzung des Korpus als Gold-Standard eine Konvertierung nötig sein. Die TÜBA-D umfasst 375.000 Token der Textsorte Dialoge (manuell annotiert). Die Annotation entspricht dem hier verwendeten Format. Allerdings verfügen die Texte über keine Interpunktion. Weiterhin steht das DEREKO-Korpus (200 Millionen Token) zur Verfügung. Dieses Korpus ist teilannotiert (Chunks, und ab 2003 topologische Felder und Teilsätze). Mit der TSNLP steht eine Testsuite zur Verfügung (Oepen et al., 1998). Diese deckt jedoch die für das beantragte Projekt entscheidenden Bereiche Verbalkomplex (inklusive Ersatzinfinitivkonstruktionen), Felderkoordination, rekursive Strukturen, Topikalisierungen und komplexe Phrasenstrukturen nicht hinreichend ab.

Als elektronisch verfügbare Wissensquellen stehen dem Projekt die Chunk-Grammatik aus der vorherigen Förderphase sowie die linguistische Datenbank CELEX (Baayen et al., 1995) und die Semantische Datenbank (SemDB) (Thielen, 1998) von Verbmobil zur Verfügung. CELEX beinhaltet Valenzrahmen für 9400 Verben. Ausführliche Angaben über obligatorische und fakultative Komplemente sind enthalten. Im Gegensatz zu CELEX enthält die SemDB für die Subkategorisierungsrahmen auch die zur jeweiligen präpositionalen Ergänzung zugehörige Präposition. Die SemDB ist jedoch ihrem Verwendungszweck entsprechend auf die Textsorte Dialoge zugeschnitten. Aus diesen Gründen soll im beantragten Projekt das IMSLex angeschafft werden. Es umfasst Subkategorisierungsrahmen für 14.000 Lemmata, darunter auch die entsprechenden Präpositionen für Präpositionalergänzungen der Verben (Bröker und Dipper, 1999).

3.3.2 Inkrementelle Annotation linguistischer Phänomenbereiche

In der ersten Projektphase sind die vom Chunkparser generierten Chunkstrukturen in einem zweiten Analyseschritt mit Hilfe des Parsers TüSBL zu kompletten Baumstrukturen vervollständigt worden.

Die hybride Parsingstrategie von TüSBL lässt sich durch das Diagramm in Abb. 1 veranschaulichen. Der Chunkparser liefert die Annotationen der drei in Abb. 1 illustrierten Schichten der PoS-, Wort- und Chunk-Ebene. Die individuellen Chunks werden von TüSBL dann zu komplexen Phrasen verbunden, mit grammatischen Funktionen etikettiert und topologischen Feldern zugeordnet.

Die bei der Annotation verwendeten Repräsentationen gehen somit vom topologischen Feldermodell des Deutschen aus, wie es Engel (1970) und Höhle (1986) zusammenfassen. Da alle syntaktischen Theorien des Deutschen die durch das Feldermodell erfassten grundlegenden topologischen Stellungsregularitäten zu behandeln haben, kann das auf dem Feldermodell basierende Annotations-Schema den Charakter einer theorieübergreifenden bzw. theorieneutralen Repräsentation für sich in Anspruch nehmen. Bei der Korpusrecherche lässt sich durch die Annotation topologischer Felder

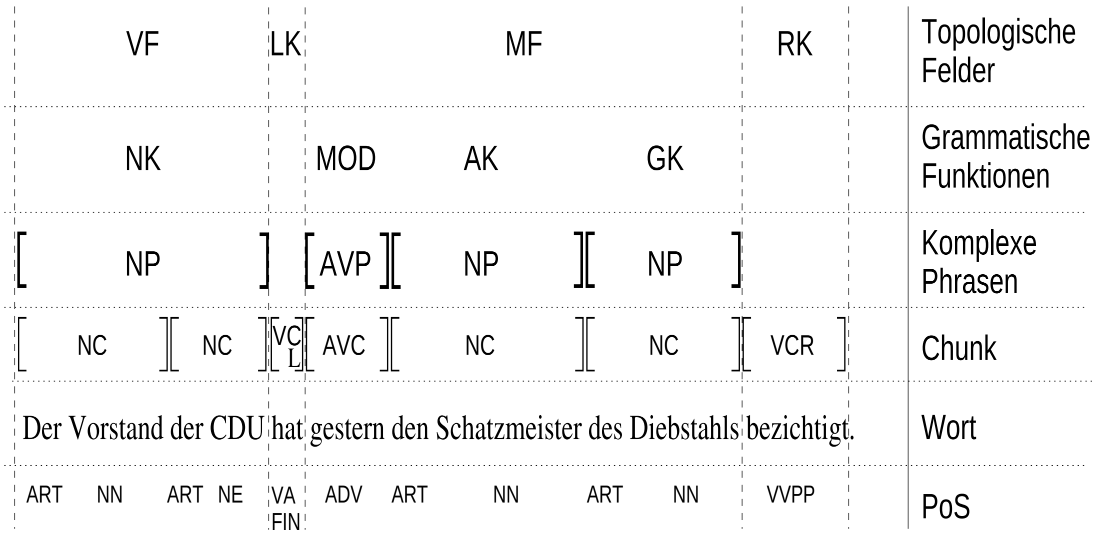

Abbildung 1: Diagramm mit Annotationsschichten

wesentlich präziser nach solchen syntaktischen Konstruktionen suchen, deren Konstituenten sich über mehrere Felder erstrecken können, wie etwa bei Partikelverben, bei partieller Vorfeldbesetzung oder bei extraponierten Relativsätzen. In der Praxis der Korpusannotation hat sich das Feldermodell zudem als robuste Repräsentation erwiesen, mittels derer in der Baumbank TüBa-D ein Korpus von 38.000 annotierten Sätzen als Trainings- bzw. Testdaten für überwachte Lernverfahren zur Verfügung steht.

Die Etikettierung komplexer Phrasen mit grammatischen Funktionen folgt der traditionellen Unterscheidung in Komplemente und Adjunkte (MOD) und ist für eine präzise Korpusrecherche einer Vielzahl von syntaktischen Konstruktionen eine wesentliche Voraussetzung. Die in A1 entwickelten syntaktischen Repräsentationen gehen somit über eine reine unetikettierte Klammeranalyse signifikant hinaus und erfüllen ein Desiderat, das bei der automatischen syntaktischen Annotation von großen Textkorpora spätestens seit Marcus et al. (1994) immer wieder eingefordert wird. Marcus et al. (1994) weisen mit Recht darauf hin, dass eine reine Klammerstruktur ohne Angabe grammatischer Funktionen wenig Aufschluss über die syntaktischen Funktionen der Einzelkonstituenten liefert. Mit Ausnahme von Parsern, die auf dem Grammatikformalismus der Dependenzgrammatik beruhen (Bröker et al., 1994; Tapanainen und Järvinen, 1997; Lesmo und Lombardo, 2000) und daher keine eigentliche Konstituentenstruktur erzeugen, und mit Ausnahme von semi-automatischen Annotationssverfahren (Brants et al., 1997) sind uns außer TüSBL keine weiteren vollautomatischen Lernverfahren zur Annotation grammatischen Funktionen bekannt.[1]

Gegenwärtig erfolgen die in Abb. 1 chunkübergreifenden syntaktischen Annotationen des Parsers TüSBL in einem einzigen Analyseschritt. In der kommenden Projektphase soll die dem Parser zugrundeliegende hybride Architektur, die gegenwärtig aus einer

[1] Siehe Ruland (2000) für einen statistischen Ansatz.

Verknüpfung von *finite-state*-Grammatiken und einem spezifischen Lernverfahren besteht, systematisch erweitert werden, indem die chunkübergreifende Annotation von grammatischen Funktionen und topologischen Feldern ebenfalls inkrementell, also in separaten Analyseschritten, erfolgen soll. Die wichtigste Motivation für die Aufgliederung in die Parse-Schritte "topologische Felder" und "grammatische Funktionen" liegt darin, dass die beiden Phänomene unterschiedlichen Ordnungsprinzipien gehorchen. Während die Struktur der topologischen Felder syntaktischen Restriktionen unterworfen ist, wird die Ordnung der grammatischen Funktionen weitgehend durch lexikalische Selektion bestimmt.

Eine derartige stärkere Inkrementalität der unterschiedlichen Analyseschritte wird auch außerhalb des Sonderforschungsbereichs in der computerlinguistischen Forschung verfolgt, v.a. in Ansätzen, die sich auf Methoden des maschinellen Lernens oder auf probabilistische Verfahren stützen, etwa von Déjean (2000); Koeling (2000); Osborne (2000); Ramshaw und Marcus (1995); Tjong Kim Sang (2000) im Bereich der Erkennung (komplexer) NPen, von Krymolowski und Dagan (2000) bei der VP-Erkennung und von Buchholz (1998); Hindle und Rooth (1993) bei der Anbindung von Präpositionalphrasen, von Fukumoto und Suzuki (1999); Karov und Edelmann (1996); Mooney (1996); Pedersen (1998); Ribas (1995); Veenstra et al. (2000) bei der Word Sense Disambiguation und von Fukumoto et al. (2000); Ge et al. (1998); Mitkov (1998, 1999) bei der Anaphoraresolution.

Zum anderen ermöglicht die Ausdifferenzierung von Analyseschritten den Einsatz unterschiedlicher Parsingverfahren in einem Shared-Task-Szenario mit vergleichender Evaluation der Ergebnisse der eingesetzten Verfahren. Derartige vergleichende Evaluationen sind in Nordamerika in den von DARPA geförderten computerlinguistischen Projekten auf den Gebieten der Spracherkennung und -analyse (ARPA, 1995; Stern, 1994) und des Message Understanding (Chinchor, 1998) seit Ende der achziger Jahre durchgeführt worden und haben sich inzwischen weltweit durchgesetzt, etwa bei den Shared-Tasks der jährlich stattfindenden Computational Natural Language Learning (CoNLL) Konferenzen (Daelemans, 2000) oder des SENSEVAL Projekts (Kilgarriff und Rosenzweig, 2000). Im Projekt A1 soll in der kommenden Phase ein derartiges Shared-Task-Szenario mit den in Abschnitt 3.3.1 genannten lern- und regelbasierten Verfahren und einer vergleichenden Evaluation der erzielten Ergebnisse einen Schwerpunkt der Projektarbeit bilden.

3.3.3 Abfrage- und Evaluationsverfahren in der Computerlinguistik

Durch die in letzter Zeit entstandenen großen syntaktisch annotierten Baumbanken ist es notwendig, geeignete Abfragewerkzeuge zur Verfügung zu haben, die es ermöglichen, die strukturellen Informationen dieser Korpora zu erschließen. Query-Tools wie z.B. *xkwic* (Christ, 1994) erlauben es, die lineare Abfolge von Sätzen oder PoS-Tags mittels regulärer Ausdrücke zu durchsuchen. Eine direkte Suche in den annotierten Strukturen, d.h. eine Suche nach bestimmten Relationen zwischen Knoten, wie z.B. Dominanz oder lineare Präzedenz, ist jedoch nicht möglich.

Selbst neuere Anfragesprachen, wie z.B. XML-QL (Deutsch et al., 1999), die sich auf die strukturellen Eigenschaften von XML-Dokumentstrukturen beziehen, lassen sich nicht direkt für die Abfrage syntaktisch annotierter Korpora verwenden. Da die meisten syntaktisch annotierten Korpora von Bäumen abweichende Datenstrukturen enthalten (z.B. überkreuzende Kanten in der NEGRA-Baumbank oder unangebundene Teilbäume in der TÜBA-D), kann die Dokumentstruktur nicht mehr direkt die syntaktische Baumstruktur widerspiegeln. Die für diese Datenstrukturen benötigten XML-Repräsentationen lassen sich nicht mehr auf intuitive Weise direkt durch XML-Anfragesprachen erschließen.

Mit dem in Stuttgart entwickelten Query-Tool TIGERSearch (Lezius und König, 2000; König und Lezius, 2000) und dem Tübinger Query-Tool VIQTORYA (Kallmeyer, 2000*c*) stehen zwei Werkzeuge zur Verfügung, die es ermöglichen, für syntaktisch annotierte Korpora komplexe strukturelle Suchanfragen zu formulieren. Die Anfragesprache von TIGER erlaubt zwar einerseits ein etwas breiteres Spektrum von Relationen zwischen Knoten, auf der anderen Seite ist TIGER jedoch restriktiver als VIQTORYA, da in TIGER keinerlei Negation erlaubt ist, während VIQTORYA atomare Negation gestattet. Im Falle der in der TÜBA-D verwendeten Datenstrukturen würde ein Verzicht auf atomare Negation eine echte Einschränkung der Ausdrucksstärke bedeuten. Im Gegensatz zu TIGER basiert die Architektur von VIQTORYA auf einer relationalen Datenbank, was eine höhere Performanz des Suchvorgangs erwarten lässt und die Verarbeitung größerer Datenmengen erlaubt.

Das Query-Tool VIQTORYA ist eine wichtige Voraussetzung für eine qualitative, phänomenorientierte Evaluation von automatischen Annotationsverfahren und soll in der kommenden Projektphase außerdem für die Bedürfnisse der phänomenorientierten Einzelprojekte erweitert werden.

Bezüglich der quantitativen Evaluation besteht das Problem der Diskrepanz zwischen der Parser-Ausgabe und dem Gold-Standard. Für einen Vergleich verschiedener Parser wird dabei überwiegend die volle Tiefe der Annotation eines Gold-Standards als Referenz genutzt. Die Güte wird mit Metriken ermittelt, die Abweichungen in den Knotenkennungen und der Klammerung von Konstituenten zwischen dem Gold-Standard und dem Ergebnis der automatischen Analyse messen. Dies gilt für Parser, die die Ausgabe einer vollen syntaktischen Analyse zum Ziel haben, genauso wie für Parser, die nur einzelne linguistische Phänomene analysieren. Die verwendete Vergleichsmetrik zwischen dem Ergebnis der automatischen Annotation und dem Gold-Standard ist überwiegend *labeled precision & recall* (z.B. Collins, 1999; Ratnaparkhi, 1998). Dabei hat sich gezeigt, dass das Trainieren probabilistischer Parser nicht unabhängig von der verwendeten Metrik ist, die zur Bewertung der automatisch erstellten Annotation benutzt wird (Goodman, 1998). Die Wahl der Metrik bestimmt also entscheidend die Ergebnisse der Annotation. Ebenso bedeutend ist die Wahl des Annotationsschemas. Die im Englischen meist als Gold-Standard verwendete Penn Treebank (Marcus et al., 1993) erlaubt z.B. keinen problemlosen Vergleich mit der Ausgabe von Parsern, die sich auf andere grammatische Theorien wie z.B. die Dependenzgrammatik stützen (Bangalore, 2000). Schließlich hat sich das meist verwendete Maß überkreuzender Kanten als wenig geeignet erwiesen, um Parser zu evaluieren, die einzelne linguistische Phänomene annotieren, wie dies im skizzierten Shared-Task-Szenario der Fall ist.

Die bekannten Lösungen für diese Probleme konzentrieren sich darauf, nur Teile der Annotation für eine Evaluation heranzuziehen (Braun, 1999; Bangalore, 2000) oder eine allgemeine Methode zu finden, mit der sich die Annotation des Gold-Standards in eine Annotation überführen lässt, die der verwendeten Annotationsmethode eher gerecht wird. Dabei wird zum einen eine weniger tiefe Annotation als vorteilhaft angesehen (Gaizauskas et al., 1998; Lin, 1998) und zum anderen ein Verfahren gesucht, das von der allgemeinsten Kopf-Beziehung ausgehend alle grammatischen Funktionen in einer Hierarchie von Dependenzbeziehungen anzuordnen versucht (Carroll et al., 1998, 1999). Das in Carroll et al. (1998) beschriebene Verfahren soll ermöglichen, Parser unterschiedlicher Sprachen zu vergleichen, darunter auch Deutsch, und ist damit neben Braun (1999) einer der wenigen Ansätze, der sich direkt auf Parser für das Deutsche bezieht.

3.4 Eigene Vorarbeiten

Zu einer Kurzfassung der Ergebnisse aus der ersten Phase siehe Abschnitt 3.3 und genauer den Arbeits- und Ergebnisbericht des Projekts. Hier werden nur diejenigen Veröffentlichungen und Arbeitsberichte von Projektangehörigen angeführt, die für Zielsetzung des Projekts relevant und im Bewilligungszeitraum 1999-2001 entstanden sind.

Publikationen und Manuskripte

Hinrichs, Erhard W. (1999): „Welchen Beitrag kann die Linguistik zu technologischen Innovationen leisten?“, *in* F. Meyer-Krahmer und S. Lange, (Hrsg.), *Geisteswissenschaften und Innovationen*, Physika Verlag, Heidelberg.

Hinrichs, Erhard W. und Tsuneko Nakazawa (1999): „Partial VP and split NP Topicalization in German: an HPSG analysis“, *in* G. Green und R. Levine, (Hrsg.), *Studies in Contemporary Phrase Structure Grammar*, Cambridge University Press, Cambridge.

Kallmeyer, Laura (1999*a*): *Tree Description Grammars and Underspecified Representations*, Dissertation, Universität Tübingen. Technischer Bericht IRCS-99-08 am Institute for Research in Cognitive Science, Philadelphia.

Kallmeyer, Laura (1999*b*): „Underspecification in Tree Description Grammars", *in* H. P. Kolb und U. Mönnich, (Hrsg.), *The Mathematics of Syntactic Structure. Trees and their Logics*, Mouton de Gruyter, S. 153–194.

Kallmeyer, Laura (2000*a*): „Komplexität von Anfragen für syntaktisch annotierte Korpora: Wh-Bewegung im Negra Korpus und in der Penn Treebank", in *Proceedings of KONVENS 2000*, S. 117–120.

Kallmeyer, Laura (2000*b*): „On the Complexity of Queries for Structurally Annotated Linguistic Data", in *Proceedings of ACIDCA'2000, Corpora and Natural Language Processing*, S. 105–110.

Kallmeyer, Laura (2000*c*): „A query tool for syntactically annotated corpora", in *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, S. 190–198.

Kallmeyer, Laura (2001): „Local Tree Description Grammars: A local extension of TAG allowing underspecified dominance relations", *Grammars*. Im Druck.

Kübler, Sandra (2001): „Braucht Nominalphrasenerkennung linguistisches Wissen?", *in* H. Lobin, (Hrsg.), *GLDV Frühjahrstagung, Universität Gießen, März 2001*, S. 225 – 233.

Kübler, Sandra und Andreas Wagner (2000): „Evaluating POS tagging under suboptimal conditions. Or: Does meticulousness pay?", in *Proceedings of ACIDCA'2000, Corpora and Natural Language Processing*.

Kübler, Sandra und Erhard W. Hinrichs (2000): „Robustes Chunkparsing mit variabler Analysetiefe", in *KONVENS 2000 / Sprachkommunikation, Ilmenau, Oktober 2000*.

Kübler, Sandra und Erhard W. Hinrichs (2001*a*): „From Chunks to Function-Argument Structure: A Similarity-Based Approach", in *Proceedings of ACL 2001, Toulouse, Juli 2001*.

Kübler, Sandra und Erhard W. Hinrichs (2001*b*): „TüSBL: A Similarity-Based Chunk Parser for Robust Syntactic Processing", in *Proceedings of HLT 2001, San Diego, Cal., März 2001*.

Müller, Frank Henrik (2001): *The German Chunk-Style-Book*. Seminar für Sprachwissenschaft, Universität Tübingen.

Müller, Frank Henrik und Tylman Ule (2001): „Satzklammer annotieren und Tags korrigieren. Ein mehrstufiges Top-Down-Bottom-Up-System zur flachen, robusten Annotierung von Sätzen im Deutschen", *in* H. Lobin, (Hrsg.), *Proceedings der GLDV-Frühjahrstagung 2001*, Gießen, S. 225–234.

Sailer, Manfred und Frank Richter (2001): „Eine XML-Kodierung für AVM-Beschreibungen", *in* H. Lobin, (Hrsg.), *Proceedings der GLDV-Frühjahrstagung 2001*, S. 161–168.

Steiner, Ilona (2000): *Eine DTD für TüBa-D Strukturen*. Unveröffentlichte Spezifikation. Universität Tübingen, SFB 441.

Steiner, Ilona (2001): „Warum "Named Entities" für die Chunk-Analyse wichtig sind", *in* H. Lobin, (Hrsg.), *GLDV Frühjahrstagung, Universität Gießen, März 2001*, S. 245 – 252.

Ule, Tylman (2000): *Extending the CES to accomodate more verbose linguistic annotation*. Talk at the Workshop "New methods and formalisms for corpus linguistics", Aix-en-Provence.

3.5 Arbeitsprogramm (Ziele, Methoden, Zeitplan)

3.5.1 Ziele

Wie in der ersten Projektphase besteht das Projektziel darin, Textkorpora des Deutschen so zu annotieren, dass die Korpusdaten für sprachwissenschaftliche Einzeluntersuchungen fruchtbar gemacht werden können.

Eine wesentliche Grundlage dafür bildet die robuste Annotation linguistischer Phänomene, die für die Makrostruktur deutscher Sätze von zentraler Bedeutung sind. Diese im Folgenden beschriebenen Phänomene sollen in der neuen Projektphase gleichrangig und gleichzeitig bearbeitet werden. Auf Basis einer Chunkanalyse soll daher untersucht werden, wie

- Chunks zu komplexen Phrasen kombiniert werden können,
- diesen komplexen Phrasen grammatische Funktionen zugewiesen werden können, und
- komplexe Phrasen in Abhängigkeit der analysierten Satztypen topologischen Feldern zuzuordnen sind.

Um eine möglichst große Annotationsqualität zu erreichen und der methodischen Vielfalt Rechnung zu tragen, die den Stand der Forschung kennzeichnet, sollen die genannten linguistischen Annotationsprobleme so weit als möglich in einem Shared-Task-Szenario durch die folgenden unterschiedlichen Annotationsverfahren bearbeitet werden:

- symbolische und subsymbolische Lernverfahren
  - *memory-based learning*
  - konnektionistische Verfahren
- regelbasierte Verfahren
  - manuell erstellte *finite-state*-Grammatiken
  - erweiterte probabilistische kontextfreie Grammatiken

Im Sinne der Vergleichbarkeit der Resultate sollen die Methoden mit gleichen Ein- und Ausgaberepräsentationen arbeiten und auf gemeinsame linguistische Datenquellen zugreifen.

Die Güte der Ergebnisse soll anhand einer quantitativen und qualitativen Evaluation systematisch überprüft werden. Hierfür soll ein Verfahren zur Evaluation partiell annotierter Strukturen entwickelt werden.

3.5.2 Methoden und Arbeitsprogramm

Wie in der ersten Projektphase soll auf Grundlage der nun vorliegenden Chunk-Analyse unter Verwendung unterschiedlicher Methoden inkrementell annotiert werden, wobei die Verfahren der Chunk-Erkennung wesentlich weiter ausgearbeitet werden sollen und Nutzen aus der Annotation weiterer linguistischer Phänomene ziehen werden.

Es sollen auch weiterhin regelbasierte Verfahren und maschinelle Lernverfahren eingesetzt werden. Bei den regelbasierten Verfahren wird zum einen versucht, auch mit *finite-state*-Methoden die chunkübergreifenden grammatischen Phänomene zu annotieren. Zum anderen sollen diese Aufgaben auch mit PCFGs modelliert werden. Neben dem bereits in der letzten Phase verwendeten symbolischen Ansatz des *memory-based learning* sollen nun auch konnektionistische Verfahren als subsymbolischer Ansatz mit einbezogen werden, da diese Verfahren unterschiedliche Extrempositionen einnehmen bezüglich der Abstraktion über Trainingsdaten. Die Shared-Tasks, die von allen genannten Verfahren bearbeitet werden sollen, betreffen die Erkennung von topologischen Feldern und grammatischen Funktionen.

Topologische Felder: Das Ziel dieser Aufgabe ist eine Einteilung von Sätzen in topologische Felder, wobei die PoS-Sequenzen der Sätze die gemeinsame Ausgangsstruktur bilden. Im folgenden Beispiel (1) ist die Abfolge der PoS-Tags für den Satz *Der Vorstand der CDU hat gestern den Schatzmeister des Diebstahls bezichtigt* dargestellt. Dabei wird den ersten vier PoS-Tags: *ART NN ART NE* das topologische Feld *Vorfeld* zugewiesen, dem nachfolgenden PoS-Tag *VAFIN* das Feld *linke Klammer (LK)* usw.

(1)	Vorfeld	LK	Mittelfeld	RK
	ART NN ART NE	VAFIN	ADV ART NN ART NN	VVPP

Grammatische Funktionen: Die gemeinsame Ausgangsrepräsentation für diese Aufgabe bildet die Annotationsebene der komplexen Phrasen[2], denen die entsprechenden grammatischen Funktionen im Satz zugewiesen werden sollen. Der Be-

[2]Diese komplexen Phrasen werden mit dem Ansatz der PCFGs erstellt.

griff “grammatische Funktion” orientiert sich somit an der Komplement-Adjunkt-Unterscheidung, wobei NP-Komplemente wie von Reis (1982) vorgeschlagen, nicht nach grammatischen Funktionen wie “Subjekt”, “Objekt” und “Indirektes Objekt”, sondern aufgrund der jeweiligen Kasusmarkierung klassifiziert werden. Im Folgenden Beispiel (2) wird der ersten Nominalphrase (NP) die grammatische Funktion Nominativkomplement (NK) zugewiesen, der zweiten NP die Funktion Akkusativkomplement (AK) und der dritten NP ein Genitivkomplement (GK). Die Adverbialphrase *gestern* wird als Modifikator (MOD) erkannt. Wird von der Chunk-Ebene als gemeinsame Ausgangsrepräsentation ausgegangen, stellt dies eine weitaus schwierigere Aufgabe dar, da manche Chunks Teil einer komplexen Phrase sind und sie außerhalb dieser Phrase keine grammatische Funktion erhalten sollen.

	NK		MOD	AK	GK	
(2)	NP		AVP	NP	NP	
	Der Vorstand d. CDU	hat	gestern	den Schatzmeister	des Diebstahls	bezichtigt.

Die wichtigste Motivation für die Aufgliederung in die Parse-Schritte “topologische Felder” und “grammatische Funktionen” liegt darin, dass die beiden Phänomene unterschiedlichen Ordnungsprinzipien gehorchen. Während die Struktur der topologischen Felder syntaktischen Restriktionen unterworfen ist, wird die Ordnung der grammatischen Funktionen weitgehend durch lexikalische Selektion bestimmt. Der konstruierte Satz *Als wir die Bedienung sahen, wollten wir unsere leeren Becher aus Plastik auf den Tisch stellen* liefert ein Beispiel. Aufgrund der bloßen Wortarteninformation ist es möglich, den Satz in Chunks, topologische Felder und Teilsätze einzuteilen. Die grammatischen Funktionen jedoch lassen sich nicht aufgrund der Wortarteninformation erkennen, sondern nur aufgrund des Valenzrahmens der entsprechenden Verben. Dieser besagt für das Verb *stellen*, dass es als Ergänzungen das Subjekt im Nominativ, das direkte Objekt im Akkusativ und eine präpositionale Ergänzung z.B. mit *auf* erfordert. Da bis zur Ebene der topologischen Felder und Teilsätze keine anderen Informationen zur Annotation als die Wortarten vorliegen müssen, ist es sinnvoll, diese Ebene in einem gesonderten Schritt zu annotieren.

Allgemein ist es zweckmäßig, Parse-Schritte in unterschiedliche Module aufzuteilen, weil dadurch die Parser-Architektur flexibler gestaltet werden kann. In der vorhergehenden Projektphase konnten so z.B. zunächst regelbasiert Chunks annotiert werden, um zur weiteren Annotation ein Lernverfahren anzuwenden. Bei einer definierten Schnittstelle ist es mit einem solchen Systemaufbau möglich, unterschiedliche Komponenten getrennt zu warten und verschiedene Methoden für verschiedene Ebenen miteinander zu kombinieren. Weiterhin kann so einfunktionalen Konstituenten wie Subordinatoren und Verben sofort ihre Funktion zugewiesen werden, während hinsichtlich ihrer syntaktischen Funktion ambigen Konstituenten wie z.B. Nominal- und Präpositional-Phrasen ihre Funktion später zugewiesen wird. Dieses Vorgehen überträgt das Konzept des *easy-first-parsing* (Abney, 1996), das bisher im Wesentlichen auf Chunks Anwendung fand, auf die Satz-Ebene.

Das hier beschriebene System integriert die syntaktische Analyse in einen Prozess inkrementeller Annotation linguistischer Datenstrukturen. Dabei werden linguistische Konstituenten für alle Verfahren in XML-Elemente gefasst und mit Attribut-Wert-Paaren versehen. Das XML-Format hat sich im Bereich der linguistischen Annotation als sehr nützlich erwiesen. So können die Module für unterschiedliche Annotationsebenen unter dem Dach eines einheitlichen Formates arbeiten und die Herkunft von verschiedenen Annotationsdaten für eine Ebene (beispielsweise bei der Verwendung verschiedener Parser für die Shared-Tasks) in der XML-Merkmalsstruktur integriert und individuell kenntlich gemacht werden (Ule, 2000). Ein weiterer Vorteil des XML-Formates besteht in der stets wachsenden Zahl der für die Verarbeitung von XML-Dokumenten verfügbaren Werkzeuge. Zudem hat sich XML als Standard zum Austausch linguistischer Daten etabliert.

Die Annotationsverfahren

Symbolische und subsymbolische Lernverfahren Bei beiden Lernansätzen, die für die Aufgaben im Shared-Task-Szenario eingesetzt werden sollen, handelt es sich um Klassifikationsverfahren, bei denen mit Hilfe von Lernalgorithmen die Eingabemuster nach Ähnlichkeiten in Klassen eingeteilt werden, mit dem Ziel, anschließend auch unbekannte ähnliche Eingabemuster korrekt klassifizieren zu können. Bei der Anwendung dieser Verfahren ist es deshalb notwendig, das linguistische Problem als Klassifizierungsaufgabe zu formulieren. Im Falle der Erkennung der topologischen Felder bedeutet das, dass jedem PoS-Tag innerhalb der PoS-Sequenz das topologische Feld zugewiesen wird, in dem das jeweilige PoS-Tag enthalten ist.

(3)	VF	VF	VF	VF	LK	…
	ART	NN	ART	NE	VAFIN	…

Das *memory-based learning* und die konnektionistischen Verfahren, die im Rahmen des Projekts eingesetzt werden, sind beide überwachte Lernverfahren. Das heißt, die Systeme lernen aus einer großen Anzahl von Trainingsbeispielen (repräsentiert durch Merkmalsvektoren), die nicht nur aus den Eingabemustern bestehen, sondern jeweils aus Paaren von Eingabemustern und korrekter Zielklasse. Beide Verfahren zeichnen sich durch die Fähigkeit aus, auch unbekannte ähnliche Eingabemuster korrekt verarbeiten zu können (Generalisierungsfähigkeit), und sind in der Lage, auch bei Subregularitäten bzw. bei unvollständiger oder fehlerhafter Eingabe gute Ergebnisse zu erzielen (Robustheit). Unvollständige oder fehlerhafte Eingabe liegt z.B. vor durch fehlende morphologische Information in den vorhandenen Baumbanken sowie durch ungrammatische Phänomene bzw. Korrekturen der gesprochenen Sprache, wie sie in der TÜBA-D vorkommen. Weiterhin betrifft dies auch Fehler in den zugrundeliegenden Annotationsebenen, wie z.B. der Tag- oder der Chunk-Ebene, falls diese Annotationen durch automatische Verfahren erstellt und nicht aus einer Baumbank extrahiert werden.

Ein wichtiges Unterscheidungsmerkmal bei maschinellen Lernverfahren ist die Art und Weise, wie die Trainingsmuster beim Lernen eingesetzt werden, um auch ungesehene Testmuster verarbeiten zu können. Die beiden Lernverfahren wurden so gewählt, dass sie bezüglich der Abstraktion über den Trainingsdaten unterschiedliche Extrempositionen einnehmen.

Beim *memory-based learning* werden die Instanzen ohne vorherigen Abstraktionsschritt unverändert gespeichert, die Generalisierung auf ungesehene Muster wird über eine Ähnlichkeitsfunktion geleistet. Dies hat sich bei der Verarbeitung natürlicher Sprache bewährt, da gerade auf diesem Gebiet nur schwer zwischen Subregularitäten in der Sprache und “Rauschen” in den Trainingsdaten unterschieden werden kann. Werden sämtliche nieder-frequenten Muster behalten, wirkt sich dies positiv bei der Verarbeitung natürlicher Sprache aus (Daelemans et al., 1999). Bei konnektionistischen Verfahren dagegen werden die Regularitäten in den Trainingsmustern durch die Konfiguration der Verbindungsgewichte repräsentiert, d.h. es wird sehr stark von den Trainingsmustern abstrahiert. Dies ist vorteilhaft, wenn die Trainingsdaten aus komplex interagierenden Merkmalen bestehen, da solche komplexe Interaktionen besser durch eine Verteilung auf verschiedene Neuronen und deren Gewichtung modelliert werden können. Dieser starke Kontrast im Abstraktionsgrad macht es besonders interessant, gerade diese beiden Verfahren zu vergleichen, da hier zwei völlig unterschiedliche Lernstrategien verfolgt werden.

Eine weitere Fragestellung ergibt sich aus den Architekturentscheidungen innerhalb der Lernverfahren, die einen entscheidenden Einfluss auf die Qualität der Klassifzierung haben können. Beim *memory-based learning* betrifft dies hauptsächlich die Wahl der Ähnlichkeitsmetrik, da auf die Abstraktion über den Trainingsinstanzen bewusst verzichtet wird. Über die Wahl der Ähnlichkeitsmetrik wird entschieden, welche Trainingsdaten zur Klassifizierung eines neuen Musters verwendet werden. Ergebnisse von Daelemans et al. (1999) haben gezeigt, dass bei linguistischen Fragestellungen die Verwendung von *information gain* als Gewichtung der Merkmale eine Verbesserung der Ergebnisse gegenüber einer Gleichbewertung der Merkmale ergibt. Es ist jedoch zu untersuchen, ob eine Gewichtung auch von den Werten der Merkmale abhängig gemacht werden sollte.

Bei konnektionistischen Verfahren besteht ein weitaus größerer Freiheitsgrad bezüglich der technischen Parameter. Es muss deshalb genau untersucht werden, welche Netzwerkarchitekturen und welche Lernalgorithmen für die Erkennung der topologischen Felder und der grammatischen Funktionen geeignet sind. Als Ausgangspunkt sollen partiell rekurrente Netzwerkarchitekturen, insbesondere Elman-Netze (Elman, 1990), verwendet werden, da bei beiden Aufgabenstellungen die lineare Abfolge von Merkmalsvektoren eine große Rolle spielt, also z.B. die Abfolge von Wörtern oder PoS-Tags in Sätzen. Ein Merkmalsvektor (eines Worts oder PoS-Tags) kann dabei nicht mehr isoliert betrachtet werden, d.h. die Klassifikation hängt nicht nur vom aktuellen Muster ab, sondern auch von den vorangegangenen Klassifikationsentscheidungen. Im Gegensatz zu vorwärtsgerichteten (*feedforward*) Netzwerken sind partiell rekurrente Netzwerkarchitekturen wie z.B. Elman-Netze jedoch in der Lage, vorangegangene Klassifikationsentscheidungen über spezielle Kontextzellen mit zu berücksichtigen. Im Gegensatz zu vollständig rekurrenten Netzen haben sie den Vorteil, dass sie mit geringfügig modifizierten Lernverfahren für *feedforward*-Netze trainiert werden können, die effizienter sind als Lernverfahren für beliebig rekurrente Netze (Rumelhart und McClelland, 1986; Zipser, 1990). Für verschiedene linguistische Aufgabenstellungen, wie z.B. Grammatikalitätsurteile, Graphem-Phonem-Zuordnungen oder das Lernen von semantischen Rollen, haben sich insbesondere Elman-Netze als geeignet erwiesen (siehe Lawrence et al. (2000), Stoianov und Nerbonne (1999), Miikkulainen und Mayberry (1999)). Für die Simulationen soll der Stuttgarter Neuronale Netze Simulator (SNNS) (Zell et al., 1993) eingesetzt werden, in dem die gängigen Netzwerkarchitekturen und Lernalgorithmen implementiert sind.

Bei der Verarbeitung von Eingaben unterschiedlicher Länge, wie sie beiden Aufgabenstellungen zugrunde liegen, ist es kaum möglich, die gesamte Sequenz (z.B. die Abfolge der PoS-Tags des ganzen Satzes) als Merkmalsvektor zu präsentieren, da Klassifikationsverfahren im Allgemeinen Merkmalsvektoren einer vorgegebenen Länge erwarten. Außerdem vergrößert sich bei zu langen Merkmalsvektoren das Problem der knappen Daten (*data sparseness*). Aus diesem Grund soll bei beiden Lernverfahren die Technik des beweglichen Fensters (*sliding window*) verwendet werden. Dabei wird statt des gesamten Satzes eine Teilfolge von *n* Mustern als Eingabe präsentiert. Dieses Teilfenster wird dann für jedes neue Muster um eine Position nach rechts verschoben. Jedes Muster wird dabei im Kontext der vorangehenden und nachfolgenden Muster betrachtet, wobei der Kontext von der Größe des gewählten Eingabefensters abhängt. Diese Technik hat den Vorteil, dass bei fest vorgegebener Größe der Vektoren dennoch der linearen Abfolge des gesamten Satzes Rechnung getragen wird. Die Beispiele (4) und (5) zeigen für den Satz in (3) die ersten beiden Trainingsmuster im Kontext der drei vorangehenden und drei nachfolgenden Merkmale (PoS-Tags). Dabei soll in (4) das erste PoS-Tag *ART* klassifiziert werden und im nachfolgenden Schritt (5) das PoS-Tag *NN*.

(4)				VF			
	–	–	–	ART	NN	ART	NE

(5)				VF			
	–	–	ART	NN	ART	NE	VAFIN

Weiterhin muss für die Bearbeitung der beiden Aufgaben genau untersucht werden, wieviele und welche linguistische Informationen in die Trainingsmuster enkodiert werden müssen, um die Aufgabe möglichst optimal lösen zu können.

Ausgegangen wird von einer Abfolge syntaktischer Kategorien, wie z.B. PoS-Tags bei der Erkennung der topologischen Felder oder der Abfolge von Phrasenkategorien (NP, PP, AVP, etc.), bzw. Chunks (NC, PC, etc.) bei der Erkennung der grammatischen Funktionen. Variiert werden sollen die folgenden Parameter:

- die *Größe des Kontexts* vor und nach der zu klassifizierenden Konstituente. Bei der Erkennung der topologischen Felder ist es z.B. wichtig, dass das Kontextfenster so gewählt ist, dass es zumindest ein verbales Element des Satzes enthält, da die *linke Klammer* und die *rechte Klammer* das Grundgerüst des Satzes darstellen.

- Information aus den darunterliegenden Annotationsebenen, wie z.B. der *lexikalische Kopf* der zu klassifizierenden Konstituente. Dies ist relevant für die Erkennung der grammatischen Funktionen, da der lexikalische Kopf darüber Aufschluss geben kann, ob die zu erkennende Konstituente vom Verb subkategorisiert wird oder nicht. Bei einem Zugriff auf tieferliegende Ebenen, wie z.B. die lexikalische Ebene oder die Ebene der Wortarten, muss abgewogen werden zwischen dem Informationsgewinn und dem Problem der *data sparseness*, das sich bei der nicht-restringierten Verwendung von lexikalischen Informationen ergeben wird.

- *Valenzinformation*, z.B. eine genauere Spezifizierung von Phrasenkategorien (insbesonderen von VPen) auf der Grundlage von Valenzinformation. Eine geeignete Klassifizierung der Verben oder auch Nomen ist eine wichtige Grundlage für die Erkennung der grammatischen Funktionen. Diese Methode hat sich z.B. auch in Lawrence et al. (2000) bei der Simulation von Grammatikalitätsurteilen bewährt. Hierfür sollen linguistische Wissensquellen, wie z.B. die SemDB oder das IMSLex, herangezogen werden.

Für die gleiche Aufgabe sollen verschiedene Merkmalssets verglichen werden. Ziel ist es, zu ermitteln, welche Arten von Information einen besonders großen Einfluss auf die Resultate aufweisen. Es soll ein ausgewogenes Verhältnis zwischen der Menge der eingehenden Informationen und der Qualität der Ergebnisse gefunden werden. Als Datenquellen für das Trainingsmaterial der beiden Lernverfahren sollen die Baumbanken TÜBA-D und NEGRA eingesetzt werden.

Regelbasierte Verfahren – *finite-state*-Formalismus und erweiterte PCFGs Neben den oben genannten Lernverfahren sollen auch regelbasierte Verfahren angewendet werden. Dabei soll einerseits ein *finite-state*-Formalismus und andererseits eine probabilistische kontextfreie Grammatik eingesetzt werden.

***Finite-State*-Formalismus** Bereits in der ersten Phase hatte sich der *finite-state*-Formalismus für die Annotation syntaktischer Phänomene bewährt. In der zweiten Phase soll die Anwendung des Formalismus über die Chunk-Ebene hinaus ausgedehnt werden. Mit den Fähigkeiten des *finite-state*-Formalismus sollen im Rahmen des projektüberspannenden Shared-Task-Szenarios sowohl topologische Felder als auch grammatische Funktionen erkannt werden. Dabei soll eine Kombination von *top-down*-Annotation und *bottom-up*-Annotation implementiert werden. Abweichend von im Allgemeinen reinen *bottom-up*-Strategien beim Chunk-Parsing werden zunächst *bottom-up* Chunks erkannt, bevor *top-down* topologische Felder und grammatische Funktionen erkannt werden. Mit Hilfe der topologischen Felder ist es möglich, Teilsätze[3] zu erkennen und so vermittels einer *divide-and-conquer*-Strategie Suchräume für die folgende Erkennung der grammatischen Funktionen einzuschränken. Durch die Iteration der Kaskaden der Grammatiken für die topologischen Felder und die Teilsätze lässt sich Rekursivität nachbilden. Dadurch lassen sich auch komplexe Sätze (mit in topologischen Feldern eingebetteten Nebensätzen) in ihrer Gesamtheit erkennen. Die so ermittelten Teilsatzgrenzen können dann auch von den Lernverfahren, die lediglich die topologischen Felder in einzelnen Haupt- und Nebensätzen erkennen können, als Eingabe für ihren Beitrag zum Shared-Task-Szenario genutzt werden.

Den nächsten Annotationsschritt im Shared-Task-Szenario stellen die grammatischen Funktionen dar. Da kein Lernverfahren angewendet wird, müssen vorhandene Wissensquellen möglichst optimal eingesetzt werden, um eine umfassende Annotation zu erreichen. Dafür sollen als linguistische Wissensquellen ein Valenzlexikon sowie morphologische Informationen und Distributionspräferenzen in den einzelnen topologischen Feldern genutzt werden. Dabei kann die Annotation jetzt auf Teilsatzgrenzen, topologischen Feldern und Chunks aufbauen. Durch diese Vorverarbeitung wird die Annotation der grammatischen Funktionen stark vereinfacht, da einerseits die Bereiche, in denen bestimmte Konstituenten auftreten können, eingegrenzt sind, und andererseits syntaktische Restriktionen und Präferenzen in den einzelnen topologischen Feldern genutzt werden können. So kann im Vorfeld eines Satzes im Deutschen typischerweise nur eine Konstituente auftreten.[4] Weitere grammatische Phänomene, die genutzt werden sollen, sind die Distributionspräferenzen im Mittelfeld, die bei morphologischer Ambiguität zur Disambiguierung dienen werden.

Für die Einbindung der Valenzinformationen ist es notwendig, Lexikoneinträge in den *finite-state*-Formalismus zu übersetzen. Da die Lexikoneinträge nicht-rekursive Strukturen darstellen, reicht die Ausdrucksstärke von regulären Ausdrücken für die Überführung von Lexikoneinträgen in endliche Automaten aus. Somit ist es möglich, die Annotation grammatischer Funktionen als einen weiteren *finite-state*-Transduktor in die Kaskade der hintereinandergeschalteten Transduktoren einzubauen. Da in Lexika meist nur eine Realisierung der Diathesen eines Verbs (z.B. der unterschiedlichen Genera Verbi) vorliegt, ist es notwendig, lexikalischen Regeln einzusetzen, die diesen Eintrag für die verschiedenen Diathesen transformieren. Meurers und Minnen (1997) zeigen, dass es möglich ist, die lexikalischen Regeln und die Interaktionen dieser Regeln in einen endlichen Automaten abzubilden. Daher lässt sich auch dieser Prozess in das Annotations-System integrieren, wobei ein Projektziel sein wird, zu untersu-

[3]Unter *Teilsatz* werden hier auch satzwertige Infinitivgruppen und einteilige Hauptsätze und damit jedes Vollverb und seine abhängigen Glieder verstanden.

[4]Für eine systematische Diskussion der wohlbekannten Ausnahmen vgl. Müller (2001).

chen, ob es sinnvoller ist, die lexikalischen Regeln **vor** der Annotation anzuwenden und somit einen Regelapparat zu erstellen, in dem die unterschiedlichen Diathesen bereits vorgesehen sind, oder ob es sinnvoller ist, einen Mechanismus zu entwickeln, der **während** der Annotation ('*on-the-fly*') solche lexikalischen Regeln anwendet. Welches Verfahren letztendlich sinnvoller ist, wird im Wesentlichen von der Anzahl der lexikalischen Regeln sowie ihres Interaktionspotenzials abhängen und daher vorrangig in einer experimentellen Untersuchung geklärt werden müssen.

Erweiterte PCFGs Zu intellektuell erstellten Regeln bietet sich als Alternative eine automatisch erstellte Grammatik an. Ähnlich der Annotation mit Hilfe von *finite-state*-Grammatiken soll auch hier mit probabilistischen kontextfreien Grammatiken ein bereits gut untersuchtes Modell zur Modellierung der Regeln verwendet werden. Dieses bietet die Möglichkeit, über die automatische Auswertung großer Datenmengen auch solche Phänomene zu annotieren, die nicht explizit in den benutzten Wissensbasen berücksichtigt werden und ist damit weit weniger abhängig vom Umfang der verfügbaren Wissensquellen als die für die gleiche Aufgabe verwendeten *finite-state*-Grammatiken.

Die Annotation grammatischer Funktionen erfolgt im Shared-Task-Szenario, also wie bei den oben genannten Lernverfahren ausgehend von Chunks morphosyntaktisch annotierter Wortformen. Als Wissensquelle steht auch diesem Ansatz ein Lexikon von Valenzdaten zur Verfügung, das Informationen zu vielen häufigen Verbformen enthält. Das Hauptinteresse liegt dabei auf der Art der Modellierung der PCFG, wobei geklärt werden soll, welche Erweiterungen einer reinen PCFG für das Parsen deutscher Sätze am geeignetsten sind und wie die vorhandenen Daten- und Wissensquellen im Modell berücksichtigt werden können, so dass sie den größten Nutzen für die Steigerung der Annotationsqualität bieten.

PCFGs sollen zuerst in ihrer reinen Form implementiert werden, die als Vergleichsbasis für die Entwicklung erweiterter Versionen dienen soll. Es hat sich gezeigt, dass PCFGs ohne Erweiterung zum einen nicht mächtig genug sind, um syntaktische Zusammenhänge innerhalb eines Satzes adäquat abzubilden, und sie zum anderen nur wenig effizient die beste Analyse aus dem gesamten Suchraum über alle syntaktischen Analysen eines Satzes wählen. Zur Eingrenzung des Suchraums lassen sich die als Eingabe bereits vorhandene Annotation von Chunks und später auch topologischer Felder nutzen. Da der Kopf des Satzes durch die Annotation der Verbklammer in der vorausgehenden Chunk-Analyse gegeben ist und im Vorfeld des Satzes in den meisten Fällen nur eine Konstituente erwartet wird, stellt die Verbindung der Chunks im Vorfeld zu einer Konstituente eine vergleichsweise einfache Aufgabe dar, die als erstes gelöst werden soll. Mit Hilfe der so eingeschränkten Ambiguität und durch Einbeziehung der in externen Wissensquellen gegebenen Valenzdaten soll erst in einem weiteren Schritt die Erkennung der Konstituenten im Mittelfeld und deren grammatischer Funktionen geleistet werden, so dass die Annotation komplexerer Strukturen auf derjenigen einfacherer Strukturen aufbaut (*easy-first parsing*) und die als Eingabe für diesen Annotationsschritt bereits annotierten Phänomene optimal genutzt werden.

Die Mächtigkeit von PCFGs soll mit Hilfe von Erweiterungen gezielt erhöht werden. Dabei sollen zunächst die *lexikalischen Köpfe* von Chunks und Phrasen berücksichtigt werden, die sich als wichtigste Erweiterung von PCFGs etabliert haben und linguistisch wohlmotiviert sind, da sie Informationen repräsentieren, die syntaktisch relevant sind, aber in reinen kontextfreien Grammatiken nicht berücksichtigt werden. Die vorhandene Annotation von Chunks ermöglicht es, diese lexikalischen Elemente aus einer bereits stark eingeschränkten Teilmenge der terminalen Elemente eines Satzes zu wählen. Die Erweiterung von PCFGs wirft dabei die wichtige Frage auf, wie die zu berücksichtigenden Informationen in das Grammatikmodell integriert werden sollen. Hierfür haben sich Modelle bewährt, die der Derivationsgeschichte eines Parse Wahrscheinlichkeiten zuordnen (*history-based models*). Der damit verbundenen Problematik der knappen Daten (*sparse data problem*) soll begegnet werden, indem man zunächst in der Derivationsgeschichte die Bildung komplexerer Strukturen von der Bildung einfacherer Strukturen abhängig macht, sich also wieder des *easy-first parsing* bedient. Die Analyse der Verbalklammer, des Vorfelds und der benutzten Valenzinformation wird so als jeweils eigener Schritt einen Teil der Derivationsgeschichte ergeben. Ähnliche Kombinationen dieser Schritte sollen dann in Äquivalenzklassen abgebildet werden, so dass für fehlende Beispiele die Häufigkeiten ähnlicher Daten herangezogen werden können.

Es wird davon ausgegangen, dass die Annotation grammatischer Funktionen mit PCFGs stark von bereits annotierten linguistischen Phänomenen profitiert. Neben der Modellierung der Derivationsgeschichte, um darin die verbleibenden Ambiguitäten in der Reihenfolge steigender Komplexität aufzulösen, bietet die robuste Annotation großer Textkorpora zunächst mit Chunks und später mit Informationen zu topologischen Feldern die Möglichkeit, für die Modellierung des probabilistischen Modells der PCFG unüberwachte Lernverfahren einzusetzen. Dazu sollen die im SFB vorhandenen Textkorpora von mehr als 200 Millionen laufender Wörter mit dem in der letzten Projektphase entwickelten Chunk-Parser automatisch annotiert werden. Die bereits teilannotierten Daten können dann ein überwacht trainiertes Modell mit Hilfe von *Bootstrapping*-Verfahren auf der großen Menge automatisch annotierter Korpora weiter verfeinern. Zum überwachten Training sollen die syntaktischen Baumbanken NEGRA bzw. TÜBA-D benutzt werden, ergänzt durch phänomenorientierte Beispielsammlungen, deren breite Abdeckung linguistischer Phänomene sie zur Repräsentation seltener Phänomene prädestiniert. Es wird erwartet, dass der Einsatz unüberwachter Lernverfahren die begrenzte Menge annotierter Daten gut ergänzen wird.

Die Annotation grammatischer Funktionen soll erfolgen, indem die Chunks der topologischen Felder zunächst zu komplexen Phrasen kombiniert werden, die zur Unterscheidung ihrer Funktion im Satz daraufhin weiter ausdifferenziert werden. Ein wichtiges Zwischenziel bei der Annotation grammatischer Funktionen ist also, anders als in den bisher beschriebenen Ansätzen, die Annotation komplexer Phrasen. Diese stellen als Konstituenten der Eingabesätze für sich eine wichtige linguistische Abstraktionsebene dar. Außerdem löst das Vorhandensein komplexer Phrasen ein wichtiges Teilproblem der Annotation, da sie dem subsymbolischen Ansatz als verbesserter Ausgangspunkt für die Annotation grammatischer Funktionen dienen.

Quantitative und qualitative Evaluation Die Szenarien der Shared-Tasks sollen helfen, die Annotation zentraler linguistischer Phänomene zu verbessern. Qualitative Evaluation ermöglicht dabei die Bewertung der Abdeckung einzelner Phänomene und ist daher besonders geeignet, die Erfüllung der Einzelaufgaben der Shared-Tasks zu überwachen. Um die qualitative Evaluation durchzuführen, soll eine Testsuite erstellt werden, die spezielle Beispiele zur Überprüfung der Phänomenbereiche "topologische Felder" und "grammatische Funktionen" enthält. Verschiedene Beispiele unterschiedlicher Komplexität für diese Phänomene können nur dann natürlichsprachliche Probleme zufriedenstellend abbilden, wenn sie sich eng an menschlichen Äußerungen orientieren. Um solche relevanten Äußerungen aus Korpora zu extrahieren und daraus die Elemente der Testsuite zu erstellen, sollen Anfragen über die Baumstruktur annotierter Korpora definiert werden, die mit Hilfe des Query-Tools VIQTORYA durchgeführt werden können. Die Anwendung von VIQTORYA auf die Phänomenbereiche der Shared-Tasks verspricht dabei Hinweise auf nützliche Erweiterungen der im Query-Tool verfügbaren Abfragelogik. So wird es wünschenswert sein, auch nach Strukturen suchen zu können, die bestimmte Informationen *nicht* enthalten, wie dies z.B. bei der Recherche von Sätzen ohne Vorfeld benötigt wird. Eine andere Erweiterung, die den Nutzen von VIQTORYA für die Recherche erhöhen wird, ist die Einbeziehung morphologischer Informationen in die Abfragen, so dass z.B. gezielt nach Konstituenten gesucht werden kann, die einen bestimmten Kasus aufweisen.

Die quantitative Evaluation zielt auf die frequenzorientierte Bewertung der Annotationsverfahren bei ihrer Anwendung auf uneingeschränkten Text. Hierfür sollen die annotierten Korpora eingesetzt werden, deren Annotation einen Gold-Standard darstellt, der automatisch mit der Ausgabe der Annotationsverfahren verglichen werden soll. Es kann aber nicht davon ausgegangen werden, dass die Beispiele für einzelne linguistische Phänomene in Korpora getrennt voneinander auftreten. Daher besteht die Notwendigkeit, nur Teile des annotierten Gold-Standards mit den Ergebnissen der Shared-Tasks zu vergleichen. Es existiert aber bisher keine befriedigende Methode, partielle und phänomenspezifische Annotation automatisch mittels eines Gold-Standards zu evaluieren, da zum einen die üblicherweise verwendeten Maße *precision*, *recall* und der *crossing brackets* nur geringe Aussagekraft bezüglich der hier zentralen Frage der Annotation einzelner Phänomene besitzen, und zum anderen bisher die Annotation grammatischer Funktionen bei der Evaluation syntaktischer Annotation nur geringe Beachtung gefunden hat. Es soll daher versucht werden, die für das Englische bekannten Verfahren der Annotation des Gold-Standards mit geringerer Informationsdichte auf das Deutsche zu übertragen und so zu erweitern, dass darin auch grammatische Funktionen berücksichtigt werden. Hierfür soll eine dependenzorientierte Annotation des Gold-Standards gewählt werden, die von gewissen Details der Konstituentenstruktur abstrahiert (Carroll et al., 1998), aber andere Informationen wie die

Existenz lexikalischer Köpfe beibehält. Auf diese Art entsteht ein *Phänokorpus* (Sailer und Richter, 2001). Es sollen also insbesondere Informationen berücksichtigt werden, wie sie auch in den automatischen Annotationsverfahren der Shared-Task "grammatische Funktionen" eine Rolle spielen, um so die Ausgabe der Annotationsverfahren mit dem entsprechend kodierten Gold-Standard gezielt im Hinblick auf die relevanten linguistischen Phänomene vergleichen zu können.

Ist die Bewertung von partieller Annotation möglich, soll auch ermittelt werden, welche Auswirkungen die Annotationsgüte eines bestimmten Phänomens auf die Güte der Annotation eines anderen Phänomens hat, also wie z.B. die Felderannotation mit hoher Güte dabei hilft, Chunks mit größerem Erfolg zu komplexen Phrasen zu verbinden. Die phänomenorientierte Evaluation ist damit gleichzeitig eine Evaluation der aufeinanderfolgenden Stufen automatischer linguistischer Annotation und soll so helfen, eine optimale Kombination von Annotationsmethoden für die einzelnen Phänomenbereiche zu finden, die in ihrer Kombination die Güte der gesamten Annotation maximiert.

3.5.3 Zeitplan

2002

- Lernverfahren
  - Empirische Untersuchung zu den einschlägigen Satztypen des Deutschen mit dem Ziel, verschiedene Merkmalssets für die Erkennung topologischer Felder zu erstellen
  - Entwicklung eines entsprechenden Trainingskorpus für die symbolischen und subsymbolischen Lernverfahren
  - Anwendung von symbolischen und subsymbolischen Lernverfahren zur Erkennung der topologischen Felder für verschiedene Architekturen und Lernalgorithmen (und für verschiedene Merkmalssets)
  - Überprüfung der Ergebnisse und gegebenenfalls Integration zusätzlicher Wissensquellen
- Regelbasierte Verfahren
  - Erstellung einer *finite-state*-Grammatik für die Erkennung von topologischen Feldern und Teilsatzgrenzen
  - automatische morphologische Annotation mit dem Werkzeug Morphy als Wissensquelle für die Disambiguierung von grammatischen Funktionen
  - Erstellung einer reinen PCFG unter Verwendung einer Baumbankgrammatik für die Annotation komplexer Phrasen im Vorfeld
  - Nutzung des Gold-Standards zur weiteren Verbesserung der Chunk-Annotation
- Evaluation und Recherche
  - Weiterentwicklung des Query-Tools VIQTORYA
  - Definition von Metriken zur Evaluation der Felderannotation und der Annotation von Teilsatzgrenzen

  - Erstellen eines Gold-Standards durch Überführung von NEGRA in eine Annäherung des TÜBA-D-Schemas
  - Erstellung einer phänomenorientierten Testsuite

2003

- Lernverfahren
  - Empirische Untersuchung zu den grammatischen Funktionen des Deutschen mit dem Ziel, verschiedene Merkmalssets zu erstellen
  - Entwicklung eines entsprechenden Trainingskorpus für die symbolischen und subsymbolischen Lernverfahren
  - Anwendung von symbolischen und subsymbolischen Lernverfahren zur Erkennung der grammatischen Funktionen ausgehend von komplexen Phrasen für verschiedene Architekturen und Lernalgorithmen (und für verschiedene Merkmalssets)

- Regelbasierte Verfahren
  - Überführung der Valenzinformationen in eine *finite-state*-Grammatik
  - Implementierung des *finite-state*-Systems zur Annotierung grammatischer Funktionen
  - Einbeziehung von lexikalischen Informationen in die PCFG
  - Erweiterung der PCFG zur Nutzung von Valenzdaten
  - Nutzung der PCFG zur Annotation komplexer Phrasen in Vorfeld und Mittelfeld

- Evaluation und Recherche
  - Weiterentwicklung des Query-Tools VIQTORYA
  - Quantitative und qualitative Evaluation für den Shared-Task "topologische Felder" und für die Annotation von Teilsatzgrenzen
  - Vergleich des Shared-Tasks "topologische Felder" in Bezug auf verschiedene Architekturen und Lernalgorithmen innerhalb der Lernverfahren und zwischen den Lernverfahren und dem *finite-state*-Ansatz
  - Evaluation des Nutzens der Annotation komplexer Phrasen als Ausgangsdaten für die anderen Annotationsverfahren
  - Definition einer Evaluationsmetrik basierend auf einem Vergleich von Dependenzen zur gezielten Evaluation des Shared-Tasks "grammatische Funktionen"

2004

- Lernverfahren
  - Überprüfung der Ergebnisse der Lernverfahren für den Shared Task "grammatische Funktionen" und gegebenenfalls Integration zusätzlicher Wissensquellen

  - Anwendung von symbolischen und subsymbolischen Lernverfahren zur Erkennung der grammatischen Funktionen ausgehend von Chunks unter Berücksichtigung der bisherigen Ergebnisse (für die modifizierte Ausgangsrepräsentation)

- Regelbasierte Verfahren
  - Verbesserung des *finite-state*-Systems zur Annotierung grammatischer Funktionen aufgrund der Ergebnisse der Evaluation
  - Nutzung der PCFG zur Annotation grammatischer Funktionen als Erweiterung der Annotation komplexer Phrasen
  - Erweitern der PCFG zur Nutzung partiell annotierter Daten mittels unüberwachtem Lernen
  - Erstellen einer Markov-Grammatik für die PCFG

- Evaluation und Recherche
  - Erstellung eines Benutzerhandbuchs für VIQTORYA
  - Quantitative und qualitative Evaluation für den Shared-Task "grammatische Funktionen"
  - Vergleich des Shared-Tasks "grammatische Funktionen" in Bezug auf verschiedene Architekturen und Lernalgorithmen innerhalb der Lernverfahren und zwischen den Lernverfahren, dem *finite-state*-Ansatz und den erweiterten PCFGs

- Annotation eines Korpus mit topologischen Feldern, komplexen Phrasen und grammatischen Funktionen durch die ermittelte optimale Kombination von Annotationsverfahren

- Abschluss offener Arbeiten und Dokumentation

3.6 Stellung innerhalb des Sonderforschungsbereichs

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

- A2 (Mönnich): Die Projekte A1 und A2 verbindet ein gemeinsames Interesse an den Konsequenzen, die die Wahl der linguistischen Datenstrukturen in Annotationssystemen hat, z.B. für die Abfragbarkeit von Korpora und damit für Anfragesprachen. Durch die in A1 geplante Weiterentwicklung der Anfragesprache in VIQTORYA ergibt sich eine enge Verbindung zu den in A2 geplanten formalen Untersuchungen von linguistischen Datenstrukturen.

- A3 (Sternefeld): Das Projekt A3 plant einen Datentypenvergleich zwischen experimentell erhobenen Daten und Korpusdaten des Deutschen, so dass A3 ein Interessent ist für die Nutzung der in A1 entwickelten syntaktisch annotierten Korpora und des Query-Tools VIQTORYA.
- A5 (Richter): Mit dem Projekt A5 gibt es durch den in A1 geplanten Einsatz der maschinellen Lernverfahren ein gemeinsames Interesse an den Methoden, die es auf der einen Seite ermöglichen, durch Generalisierung über die Endlichkeit von Korpusdaten hinauszugehen, und die auf der anderen Seite eine Zusammenführung von Regularitäten und Subregularitäten ermöglichen. Eine weitere Verbindung zum Projekt A5 besteht darin, dass zur Erschließung unikaler Elemente die in A1 entwickelten korpuslinguistischen Methoden angewandt werden.
- B1 (Berger), B8 (Raecke), B11 (Butzenberger): In diesen Projekten sollen Korpora annotiert 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): Das Projekt B3 beschäftigt sich ebenso wie A1 schwerpunktmäßig mit deutschen Daten. Dadurch ist B3 im Bereich der Korpusarbeit ein Interessent für die in A1 entwickelten Verfahren zur Erschließung deutscher Korpora. Andererseits sind durch die phänomenzentrierten Untersuchungen in B3 wichtige Impulse für die von A1 entwickelten Repräsentations- und Erschließungssysteme zu erwarten.
- B6 (Koch): Das Projekt A1 plant für die automatische Annotation von Textkorpora die Einbindung lexikalischer Ressourcen, was u.a. eine Klassifizierung lexikalischer Elemente notwendig machen wird. Hierfür kann das in Projekt B6 bereitstehende Beschreibungssystem und die damit verbundene Erfahrung mit lexikalischen Ressourcen Basis für einen fruchtbaren Austausch bieten.
- B10 (von Stechow): Durch die im Projekt B10 geplante Weiterentwicklung einer Datenbank ist B10 ein Interessent für die Nutzung der in A1 entwickelten syntaktisch annotierten Korpora und des Recherchewerkzeugs VIQTORYA. Der Austausch über Annotationsstandards soll weitergeführt werden und wichtige Impulse für die Erweiterung von VIQTORYA im Hinblick auf die Abfragbarkeit von Informationen, die über reine Baumstrukturen hinausgehen, sind zu erwarten.
- B13 (Winkler): Mit dem Projekt B13 ist eine Zusammenarbeit im Bereich Korpusrecherche von Ellipsen geplant. Elliptische Strukturen stellen eine Robustheitsprüfung des in A1 verwendeten Annotationsschemas dar. Untersucht werden sollen auch die Implikationen, die sich für die Ausdrucksstärke der in VIQTORYA definierten Anfragesprache ergeben.

- C1 (Reis/Hinrichs): In der kommenden Phase soll die Weiterentwicklung des Query-Tools VIQTORYA in enger Kooperation mit dem Projekt C1 und orientiert an den Bedürfnisse der phänomenorientierten Einzelprojekte des SFB erfolgen. Wie auch in der letzten Phase soll weiterhin ein regelmäßiger Wissensaustausch mit C1 stattfinden.

Organisation einer Sommerschule In Anbindung an den SFB 441 soll im Sommer 2002 die Sommerschule "Empirische Sprachwissenschaft und Maschinelle Sprachverarbeitung" stattfinden. Sie ist eine Kooperation der Projekte A1 (Hinrichs) und B1 (Berger) und soll Doktoranden aus Osteuropa die Möglichkeit geben, sich intensiv mit der Korpuslinguistik zu beschäftigen und Kontakt zu Lehrenden sowohl aus Deutschland als auch aus osteuropäischen Ländern herzustellen. Die Anbindung an den SFB 441 macht es möglich, theoretische Untersuchungen zentraler linguistischer Phänomene in slavischen Sprachen und die Demonstration verschiedener korpuslinguistischer Annotations- und Suchtools zu verbinden. Für die beiden federführenden Projekte soll durch die Sommerschule eine Intensivierung der Zusammenarbeit auf dem Gebiet der automatischen Annotation erreicht werden, wobei vor allem die Charakteristika der slavischen Sprachen berücksichtigt werden sollen. Für den gesamten SFB ergibt sich die Möglichkeit, von einem Erfahrungsaustauch mit den Lehrenden und Teilnehmern aus slavischen Ländern zu profitieren.

Verbindung zu Forschergruppen außerhalb Tübingens Neben der Zusammenarbeit mit anderen Teilprojekten des SFB 441 sollen Kooperationen mit in- und ausländischen Forschergruppen etabliert bzw. fortgesetzt werden, die ebenfalls auf dem Gebiet der Computerlinguistik und der Texttechnologie arbeiten:

- Arbeitsbereich Angewandte Sprachwissenschaft und Computerlinguistik, Institut für deutsche Sprache und mittelalterliche Literatur, Universität Gießen: Im beantragten Verbundprojekt "Medienintensive Lehrmodule in der Computerlinguistik-Ausbildung" (MiLCA) soll ein intensiver Austausch über Texttechnologie in der Korpuslinguistik mit Prof. Dr. Henning Lobin stattfinden. Hierbei soll vor allem die Definition von XML-Strukturen im Mittelpunkt stehen.

- AT&T Laboratories, New Jersey: Die Koperation mit Dr. Steven Abney, einem ehemaligen Mitarbeiter am Lehrstuhl Allgemeine Sprachwissenschaft / Computerlinguistik, soll in den Bereichen des Chunkparsing und der statistischen Methoden in der Computerlinguistik, vor allem bezüglich *Bootstrapping*, fortgesetzt werden.

- Center for Dutch Language and Speech, Antwerpen, und Induction of Linguistic Knowledge am Center for Language Studies, Tilburg: Der bereits bestehen-de Kontakt zu Prof. Dr. Walter Daelemans auf dem Gebiet des *memory-based learning* soll in der nächsten Phase fortgesetzt werden.

- Institut für Deutsche Sprache, Mannheim, und Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart: Im Rahmen des Projekts "Deutsches Referenzkorpus" soll die Kooperation auf den Gebieten der Erstellung, der partiellen inkrementellen Annotation und der Recherche auf Korpora fortgeführt werden.

- Xerox Research Center Europe (XRCE), Grenoble: Es wird eine Kooperation mit Jean-Pierre Chanod, dem Leiter der Forschungsgruppe zu Parsing und Semantik, angestrebt. Im Rahmen dieser Kooperation soll der *Xerox Incremental Parser* für das Deutsche adaptiert werden.

Zitierte Literatur

Abney, Steven (1996): „Partial Parsing via Finite-State Cascades", in *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Aït-Mokhtar, Salah und Jean-Pierre Chanod (1997): „Incremental Finite-State Parsing", in *Proceedings of ANLP 97*, Washington, S. 72–79.

Argamon, Shlomo, Ido Dagan und Yuval Krymolowski (1998): „A Memory-Based Approach to Learning Shallow Natural Language Patterns", in *Proceedings of the COLING/ACL 1998, Montreal, Canada*, S. 67 – 73.

ARPA (Hrsg.) (1995): *Proceedings of the ARPA Spoken Language Systems Technology Workshop, Austin, Texas*, Morgan Kaufman: San Francisco.

Baayen, R. H., R. Piepenbrock und L. Gulikers (1995): *The CELEX Lexical Database*, Technischer bericht, Linguistic Data Consortium, University of Pennsylvania.

Bangalore, Srinivas (2000): *Performance Evaluation of SuperTagging for Partial Parsing*, Band 16 in *Text, Speech and Language Technology*, Kluwer Academic Publishers, Kapitel 11, S. 203–220.

Beil, Franz, Glenn Carroll, Detlef Prescher, Stefan Riezler und Mats Rooth (1999): „Inside-Outside Estimation of a Lexicalized PCFG for German", in *Proceedings of the 37th Annual Meeting of the ACL*, Maryland.

Bos, Johan (1995): „Predicate Logic Unplugged", in *Proceedings of the 10th Amsterdam Colloquium*.

Brants, Thorsten (1999): *Tagging and Parsing with Cascaded Markov Models*, Dissertation, Universität des Saarlandes, Saarbrücken.

Brants, Thorsten, Wojiech Skut und Brigitte Krenn (1997): „Tagging Grammatical Functions", in *Proceedings of EMNLP-2 1997, Providence, RI.*

Braun, Christian (1999): *Flaches und robustes Parsen deutscher Satzgefüge*, Diplomarbeit, Lehrstuhl Prof. Dr. Hans Uszkoreit, Universität des Saarlandes, Saarbrücken.

Bröker, Norbert und Stefanie Dipper (1999): „Zur Konstruktion von Lexika für die maschinelle syntaktische Analyse", *in* J. Gippert und P. Olivier (Hrsg.), *Multilin-*

guale Corpora - Codierung, Strukturierung, Analyse. 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung, Enigma corporation, Prag.

Bröker, Norbert, Udo Hahn und Susanne Schacht (1994): „Concurrent lexicalized dependency parsing: the ParseTalk model“, in *Proceedings of COLING 94, Kyoto, Japan*.

Buchholz, Sabine (1998): „Distinguishing Complements from Adjuncts Using Memory-Based Learning“, in *Proceedings of the ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*, S. 41 – 48.

Buchholz, Sabine, Jorn Veenstra und Walter Daelemans (1999): „Cascaded Grammatical Relation Assignment“, in *Proceedings of EMNLP/VLC-99, University of Maryland, USA, June 21-22, 1999*, S. 239 – 246.

Cardie, Claire und David Pierce (1998): „Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification“, in *Proceedings of ACL-Coling*, S. 218 – 224.

Carroll, John, Guido Minnen und Ted Briscoe (1999): „Corpus annotation for parser evaluation“, in *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora (LINC-99)*, Bergen, Norwegen, S. 35–41.

Carroll, John, Ted Briscoe und Antonio Sanfilippo (1998): „Parser evaluation: a survey and a new proposal“, in *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spanien, S. 447–454.

Charniak, Eugene (1997): „Statistical Parsing with a Context-free Grammar and Word Statistics“, in *Proceedings of the 14th AAAI*, Menlo Park.

Charniak, Eugene (1999): *A Maximum-Entropy-Inspired Parser*, Technischer Bericht CS-99-12, Brown University.

Chinchor, Nancy (1998): „MUC-7 Named Entity Task Definition“, in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Christ, Oliver (1994): „A modular and flexible architecture for an integrated corpus query system“, in *Proceedings of COMPLEX'94*.

Clematide, Simon und Martin Volk (2001): „Linguistische und semantische Annotation eines Zeitungskorpus“, *in* H. Lobin (Hrsg.), *Proceedings der GLDV-Frühjahrstagung 2001*, Gesellschaft für Linguistische Datenverarbeitung, Gießen, S. 201–209.

Collins, Michael (1999): *Head-Driven Statistical Models for Natural Language Parsing*, Dissertation, University of Pennsylvania.

Daelemans, Walter (Hrsg.) (2000): *Proceedings of CoNLL-2000: International Conference on Computational Natural Language Learning*, Lissabon.

Daelemans, Walter, Jakub Zavrel und Antal van den Bosch (1999): „Forgetting Exceptions is Harmful in Language Learning“, *Machine Learning: Special Issue on Natural Language Learning* **34**, S. 11–43.

Déjean, Hervé (2000): „ALLiS: a Symbolic Learning System for Natural Language Learning“, in *Proceedings of CoNLL-2000, Lissabon, Portugal*.

Deutsch, Alin, Mary Fernandez, Daniela Florescu, Alon Levy und Dan Suciu (1999): „A Query Language for XML“, in *Proceedings of the International World Wide Web Conference (WWW)*, Band 31, S. 1155–1169.

Eisner, Jason (2000): *Bilexical Grammars and Their Cubic-Time Parsing Algorithms*, Band 16 in *Text, Speech and Language Technology*, Kluwer Academic Publishers, Dordrecht, Kapitel 3.

Elman, Jeffrey L. (1990): „Finding Structure in Time“, *Cognitive Science* **14**, S. 179–211.

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

Finch, Steve und Andrei Mikheev (1997): „A Workbench for Finding Structure in Texts“, in *Applied Natural Language Processing*, Washington D.C.

Fukumoto, Fumiyo, Hiroyasu Yamada und Ruslan Mitkov (2000): „Resolving Overt Pronouns in Japanese using Hierarchical VP Structures“, in *Proceedings of the International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications, ACIDCA 2000, Monastir, Tunesien*, S. 152 – 157.

Fukumoto, Fumiyo und Yoshimi Suzuki (1999): „Word Sense Disambiguation in Untagged Text Based on Term Weight Learning“, in *Proceedings of EACL ’99, Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norwegen, S. 209 – 216.

Gaizauskas, R., M. Hepple und C. Huyck (1998): „Modifying existing annotated corpora for general comparative evaluation of parsing“, in *Proceedings of the LRE Workshop on Evaluation of Parsing Systems*, Granada, Spanien.

Ge, Niyu, John Hale und Eugene Charniak (1998): „A statistical approach to anaphora resolution“, in *Proceedings of the Sixth Workshop on Very Large Corpora*.

Goodman, Joshua (1998): *Parsing Inside-Out*, Dissertation, Harvard University.

Hindle, Donald und Mats Rooth (1993): „Structural Ambiguity and Lexical Relations“, *Computational Linguistics* **19** (1), S. 103 – 120.

Hinrichs, Erhard W., Sandra Kübler, Valia Kordoni und Frank H. Müller (2000): „Robust Chunk Parsing for Spontaneous Speech“, *in* W. Wahlster (Hrsg.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, Springer Verlag, Berlin, S. 163–182.

Höhle, Tilman 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*, Band 3, S. 329–340.

Karov, Yael und Shimon Edelmann (1996): „Learning Similarity-Based Word Sense Disambiguation from Sparse Data“, in *Proceedings of the Fourth Workshop on Very Large Corpora*, S. 42–55.

Kilgarriff, Adam und Joseph Rosenzweig (2000): „English Senseval: Report and Results“, in *Proceedings of LREC 2000, Athen, Griechenland*.

Koeling, Rob (2000): „Chunking with Maximum Entropy Models“, in *Proceedings of CoNLL-2000, Lissabon, Portugal*.

König, Esther und Wolfgang Lezius (2000): „A description language for syntactically annotated corpora“, in *Proceedings COLING 2000*, S. 1056–1060.

Krymolowski, Yuval und Ido Dagan (2000): „Incorporating Compositional Evidence in Memory-Based Partial Parsing“, in *Proceedings of the ACL 2000, Hong Kong*.

Lawrence, Steve, C. Lee Giles und Sandiway Fong (2000): „Natural Language Grammatical Inference with Recurrent Neural Networks“, *IEEE Transactions on Knowledge and Data Engineering* **12** (1), S. 126–140.

Lesmo, Leonardo und Vincenzo Lombardo (2000): „Automatic assignment of grammatical relations“, in *Proceedings of LREC 2000, Athen, Griechenland*.

Lezius, Wolfgang und Esther König (2000): „Towards a Search Enginge for Syntactically Annotated Corpora“, in *5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2000), Ilmenau*, S. 113 – 116.

Lin, Dekang (1998): „A Dependency-based Method for Evaluating Broad-Coverage Parsers“, *Natural Language Engineering* **4** (2), S. 97–114.

Marcus, M., S. Santorini und M. Marcinkiewicz (1993): „Building a large annotated corpus of English: the Penn Treebank“, *Computational Linguistics* **19** (2), S. 313–330.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Anne Bies, Mark Ferguson, Karen Katz und Britta Schasberger (1994): „The Penn Treebank: Annotating Predicate Argument Structure“, in *Proceedings of HLT 94, Plainsboro, New Jersey*.

Meurers, W. Detmar und Guido Minnen (1997): „A Computational Treatment of Lexical Rules in HPSG as Covariation in Lexical Entries“, *Computational Linguistics* **23** (4), S. 543–568.

Miikkulainen, Risto und M. R. III Mayberry (1999): „Disambiguation and Grammar as Emergent Soft Constraints“, *in* B. J. MacWhinney (Hrsg.), *Emergentist Approaches to Language*, Erlbaum, Hillsdale, NJ.

Mitkov, Ruslan (1998): „Robust pronoun resolution with limited knowledge“, in *Proceedings of COLING'98/ACL'98, Montreal, Canada*.

Mitkov, Ruslan (1999): „Towards automatic annotation of anaphoric links in corpora“, *International Journal of Corpus Linguistics* **4** (2).

Mooney, Raymond J. (1996): „Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning“, in *Proceedings of the 1st Conference on Empirical Methods in NLP, Philadelphia, PA*, S. 82–91.

Müller, Stefan (2001): *Complex Predicates: Verbal Complexes, Resultative Constructions, and Particle Verbs in German*, Habilitationsschrift, Universität des Saarlandes, Saarbrücken.

Muñoz, Marcia, Vasin Punyakanok, Dan Roth und Dav Zimak (1999): „A learning approach to shallow parsing“, in *Proceedings of EMNLP-WVLC'99*, Association for Computational Linguistics.

Neumann, Günter und Jakub Piskorski (2001): *A Shallow Text Processing Core Engine*. Under review by the Journal of Computational Intelligence.

Oepen, Stephan, Klaus Netter und Judith Klein (1998): „TSNLP – Test Suites for Natural Language Processing“, *in* J. Nerbonne (Hrsg.), *Linguistic Databases*, Band 77 in *CSLI Lecture Notes*, CSLI, Stanford.

Osborne, Miles (2000): „Shallow Parsing as Part-of-Speech Tagging“, in *Proceedings of CoNLL-2000, Lissabon, Portugal*.

Pedersen, Ted (1998): *Learning Probabilistic Models of Word Sense Disambiguation*, Dissertation, School of Engineering and Applied Science,Southern Methodist University, Dallas, TX.

Pereira, Fernando und Yves Schabes (1992): „Inside-outside reestimation from partially bracketed corpora“, in *Proceeding of ACL'92*, Newark, Delaware, S. 128–135.

Pinkal, Manfred (1996): „Radical underspecification“, *in* P. Dekker und M. Stokhof (Hrsg.), *Proceedings of the 10th Amsterdam Colloquium*, Universität Amsterdam.

Plaehn, Oliver (1999): *Probabilistic Parsing with Discontinuous Phrase Structure Grammar*, Magisterarbeit, Universität des Saarlandes.

Ramshaw, Lance A. und Mitchell P. Marcus (1995): „Text Chunking Using Transformation-Based Learning“, in *Proceedings of the ACL Third Workshop on Very Large Corpora*, S. 82 – 94.

Ratnaparkhi, Adwait (1998): *Maximum Entropy Models for Natural Language Ambiguity Resolution*, Dissertation, University of Pennsylvania.

Ratnaparkhi, Adwait (1999): „Learning to Parse Natural Language with Maximum Entropy Models“, *Machine Learning* **34**, S. 151–175.

Reis, Marga (1982): „Zum Subjektbegriff im Deutschen“, *in* W. Abraham (Hrsg.), *Satzglieder im Deutschen: Vorschläge zur syntaktischen, semantischen und pragmatischen Fundierung*, Tübingen, S. 171 – 211.

Reyle, Uwe (1993): „Dealing with Ambiguities by Underspecification: Construction, Representation and Deduction“, *Journal of Semantics* **10** (2), S. 123–179.

Ribas, Francesc (1995): „On Learning More Appropriate Selectional Restrictions“, in *7th Conference of the European Chapter of the ACL*, S. 112–118.

Richter, Frank und Manfred Sailer (1999): „Underspecified Semantics in HPSG“, *in* H. Bunt und R. Muskens (Hrsg.), *Computing Meaning*, Studies in Linguistics and Philosophy, Kluwer Academic Publishers, S. 95–112.

Roth, Dan und Dmitry Zelenko (1998): „Part of Speech Tagging Using a Network of Linear Separators“, in *COLING-ACL*, S. 1136–1142.

Ruland, Tobias (2000): „A Context-Sensitive Model for Probabilistic LR Parsing of Spoken Language with Transformation-Based Postprocessing“, in *Proceedings of COLING 2000, Saarbrücken*.

Rumelhart, David E. und James L. McClelland (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Band 1: Foundations, The

MIT Press.

Sailer, Manfred und Frank Richter (2001): „Eine XML-Kodierung für AVM-Beschreibungen“, *in* H. Lobin (Hrsg.), *Proceedings der GLDV-Frühjahrstagung, Gießen*, S. 161 – 167.

Schmid, Helmut und Sabine Schulte im Walde (2000): „Robust German Noun Chunking with a Probabilistic Context-Free Grammar“, in *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, Saarbrücken/Nancy/Luxembourg, S. 726–732.

Skut, Wojciech und Thorsten Brants (1998): „A Maximum-Entropy Partial Parser for Unrestricted Text“, in *Sixth Workshop on Very Large Corpora (WVLC-6)*, Montréal, Québec, Canada.

Steiner, Ilona (1995): *Verschiedene Varianten von Backpropagation angewandt auf NETtalk*. Studienarbeit, Wilhelm-Schickard-Institut für Informatik, Universität Tübingen.

Stern, Richard (Hrsg.) (1994): *Spoken language technology: Proceedings of a Workshop held at Plainsboro, New Jersey*.

Stoianov, Ivelin und John Nerbonne (1999): „Connectionist grapheme to phoneme conversion: exploring distributed representation“, in *Proceedings of CLIN ‘99*, S. 181–192.

Tapanainen, Pasi und Timo Järvinen (1997): „A non-projective dependency parser“, in *Proceedings of ANLP’97, Washington, D.C.*

Thielen, Christine (1998): *Verbrahmen-Kodierungen in den semantischen Datenbanken*, Memo 135, DFKI Saarbrücken, Saarbrücken.

Tjong Kim Sang, Erik (2000): „Text Chunking by System Combination“, in *Proceedings of CoNLL-2000, Lissabon, Portugal*.

Veenstra, Jorn (1998): „Fast NP Chunking Using Memory-Based Learning Techniques“, *in* F. Verdenius und W. van den Broek (Hrsg.), *Proceedings of Benelearn*, Wageningen, Niederlande, S. 71 – 79.

Veenstra, Jorn, Antal van den Bosch, Sabine Buchholz, Walter Daelemans und Jakub Zavrel (2000): „Memory-Based Word Sense Disambiguation“, *Computers and the Humanities, Special Issue on Senseval, Word Sense Disambiguations* **34**, S. 1 – 2.

Wauschkuhn, Oliver (1999): *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora*, Berichte aus der Informatik, Shaker Verlag, Aachen. Zugl.: Stuttgart, Univ., Diss., 1999.

Zell, Andreas (1994): *Simulation Neuronaler Netze*, Addison-Wesley (Deutschland).

Zell, Andreas, Niels Mache, Ralf Hübner, Günter Mamier, Michael Vogt, Michael Schmalzl und Kai-Uwe Herrmann (1993): *Neural Network Simulation Environments*, Kluwer Publishers, Kapitel: SNNS (Stuttgart Neural Network Simulator).

Zipser, David (1990): „Subgrouping reduces complexity and speeds up learning in recurrent networks“, *in* T. S. Touretzky (Hrsg.), *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann Publishers, Inc., S. 638–641.

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	1,5	152.400	BAT IIa/2	3	144.000	BAT IIa/2	3	144.000	BAT IIa/2	3	144.000
	Wiss.HK	1	33.600	Wiss.HK	1	33.600	Wiss.HK	1	33.600	Wiss.HK	1	33.600
	zusammen		186.000	zusammen		177.600	zusammen		177.600	zusammen		177.600
SV				Kostenkategorie oder Kennziffer		Betrag in DM	Kostenkategorie oder Kennziffer		Betrag in DM	Kostenkategorie oder Kennziffer		Betrag in DM
				515		6.000	515		0	515		0
				522		14.700	522		4.500	522		4.500
				zusammen		20.700	zusammen		4.500	zusammen		4.500
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 Teilprojekt in Std./Woche (beratend: B)**	**auf dieser Stelle im SFB tätig seit**	**beantragte Einstufung in BAT**
Grundausstattung						
3.7.1.1	1. Hinrichs, Erhard, Prof. Dr.	Syntax, Semantik, Comp.-Ling.	SfS	5		
wissenschaftl.	2. Gerdemann, Dale, Dr.	Comp.-Ling. finite-state	SfS	5		
Mitarbeiter	3. Kübler, Sandra, M.A.	Comp.-Ling. MBL	SfS	5		
(einschl. HK)	4. Saile, Jochen, M.A.	Systemprog.	SfS	4		
3.7.1.2 nichtwiss. Mitarbeiter	Stoll, Cornelia	Sekretariat	SfS	5		
Ergänzungsausstattung						
3.7.1.3	5. Steiner, Ilona, M.A.	Konnektionismus		19,25	1.4.00	BAT IIa/2
wissenschaftl.	6. Müller, Frank, 1. Staatsex.	finite-state		19,25		BAT IIa/2
Mitarbeiter	7. Ule, Tylman, M.A.	stat. Parsing		19,25		BAT IIa/2
(einschl. HK)	8. Wunsch, Holger	Comp.-Ling.		20		wiss. HK
3.7.1.4 nichtwiss. Mitarbeiter						

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

Aufgabenbeschreibung von Mitarbeitern der Grundausstattung

1. Leitung des Teilprojekts und Betreuung der damit verbundenen Forschungsarbeiten; Lehrtätigkeit, Betreuung und Anleitung von Abschlussarbeiten.

2. Beratung in computerlinguistischen Aspekten, vor allem auf dem Gebiet des *finite-state*-Formalismus; Lehrtätigkeit, Mitbetreuung von Abschlussarbeiten.

3. Beratung in computerlinguistischen Aspekten, Mitarbeit auf dem Gebiet des *memory-based learning*; Lehrtätigkeit.

4. Systembetreuung, Wartung und Akquisition von Software.

Aufgabenbeschreibung von Mitarbeitern der Ergänzungsausstattung

5. Stellenbeschreibung: Die Stelle soll mit einer Computerlingistin/einem Computerlinguisten besetzt werden, der/die in den Projektbereichen "Maschinelle Lernverfahren" und "Evaluation und Recherche" mitarbeitet. Hierfür sind Erfahrung im Umgang mit Korpora, Taggern, Suchwerkzeugen und maschinellen Lernverfahren notwendig.

   Es ist vorgesehen, diese Stelle mit Frau Ilona Steiner zu besetzen. Frau Steiner hat an der Entwicklung des Query-Tools VIQTORYA mitgearbeitet. Sie soll das Query-Tool in der beantragten Förderphase weiterentwickeln und in die geplante Datenevaluation einbringen. Im Bereich "Maschinelle Lernverfahren" soll Frau Steiner die Untersuchungen zu den Shared-Tasks "Topologische Feldererkennung" und "Annotation grammatischer Funktionen" mit subsymbolischen Lernverfahren durchführen und vergleichend evaluieren.

   Frau Steiner verfügt über exzellente Kenntnisse hinsichtlich der theoretischen Grundlagen und des experimentellen Einsatzes konnektionistischer Lernverfahren. Sie hat eine Studienarbeit über das Thema "Verschiedene Varianten von Backpropagation angewandt auf NETtalk" verfasst. Ihr Dissertationsvorhaben "Konnektionistische Modellierung und Simulation grammatischer Phänomene" steht in unmittelbarem Zusammenhang mit den geplanten Projektarbeiten.

6. Stellenbeschreibung: Die Stelle soll mit einer Computerlingistin/einem Computerlinguisten besetzt werden, der/die in den Projektbereichen "Chunkparsing und *finite-state*-Methoden" und "Evaluation und Recherche" mitarbeitet. Hierfür sind Erfahrung im Umgang mit Korpora, Taggern, *finite-state*-Grammatiken sowie Kenntnisse auf den Gebieten der Syntax des Deutschen und der Computerlexikographie notwendig.

   Es ist vorgesehen, diese Stelle mit Herrn Frank Müller zu besetzen. Im Bereich "Regelbasierte Verfahren" soll Herr Müller die Untersuchungen zu den Shared-

Tasks “Topologische Feldererkennung” und “Annotation grammatischer Funktionen” mit *finite-state*-Grammatiken durchführen und vergleichend evaluieren.

Herr Müller verfügt über exzellente Kenntnisse hinsichtlich der theoretischen Grundlagen und der praktischen Implementierung von *finite-state*-Grammatiken. Er hat die in A1 verwendete Chunkgrammatik implementiert und soll sie in der kommenden Projektphase für chunkübergreifende Strukturen erweitern. Sein Dissertationsvorhaben “Nutzung linguistischer Wissensquellen zur automatischen, robusten Annotation von unrestringierten Textkorpora des Deutschen” steht in unmittelbarem Zusammenhang mit den geplanten Projektarbeiten.

7. Stellenbeschreibung: Die Stelle soll mit einer Computerlingistin/einem Computerlinguisten besetzt werden, der/die in den Projektbereichen “Erweiterte probabilistische kontextfreie Grammatiken” und “Evaluation und Recherche” mitbearbeitet. Hierfür sind Erfahrung im Umgang mit Korpora, Taggern und überwachten Lernverfahren sowie Kenntnisse über statistische Verfahren in der Computerlinguistik und über Grammatikformalismen für natürliche Sprachen notwendig.

   Es ist vorgesehen, diese Stelle mit Herrn Tylman Ule zu besetzen. Im Bereich “Regelbasierte Verfahren” soll Herr Ule die Untersuchungen zum Shared-Task “Annotation grammatischer Funktionen” und zur Erkennung komplexer Phrasen mit erweiterten probabilistischen kontexfreien Grammatiken durchführen und vergleichend evaluieren.

   Herr Ule besitzt ein fundiertes theoretischen Wissen über den Einsatz statistischer Verfahren in der Computerlinguistik. Sein Dissertationsvorhaben “Parsing syntaktischer Strukturen des Deutschen mit erweiterten PCFGs” steht in unmittelbarem Zusammenhang mit den geplanten Projektarbeiten. Er verfügt über praktische Erfahrungen bei der computerlinguistischen Aufarbeitung und XML-Kodierung großer Textmengen sowie ihrer automatischen Annotation.

8. Zu besetzen durch eine(n) fortgeschrittene(n) Computerlinguistikstudentin/en.

   Aufgabenbereich: Unterstützung bei der Weiterentwicklung des Query-Tools VIQTORYA, vor allem Übernahme der Implementierungsaufgaben, die sich im Rahmen der Erweiterung der Query-Sprache und der Aufnahme neuer Informationsarten in die zugrundeliegende Datenbank ergeben.

   Voraussetzung: Für diese Aufgabe ist Erfahrung auf dem Gebiet der Implementierung großer Projekte nötig. Es werden Programmierkenntnisse in Java und SQL-Kenntnisse vorausgesetzt.

   Es ist vorgesehen, diese Stelle mit Herrn Holger Wunsch zu besetzen. Zu Projektbeginn wird Herr Wunsch sein Studium mit dem Grad eines M.A. abgeschlossen haben.

3.7.2 Aufgliederung und Begründung der Sächlichen Verwaltungsausgaben (nach Haushaltsjahren)

An Kleingeräten sind zwei Rechnerarbeitsplätze (Pentium II PCs) im Wert von insgesamt 8 000 DM vorhanden, die ausschließlich den Projektmitarbeitern zur Verfügung stehen. Beide PCs laufen unter dem Betriebssystem Linux. Das Projekt kann außerdem, wie auch schon in der ersten Phase, auf die am Lehrstuhl Allgemeine Sprachwissenschaft/Computerlinguistik vorhandene Rechnerausstattung zurückgreifen. Für besonders rechenintensive Arbeiten steht ein Sun Ultra 450 Enterprise Server im Wert von 65 000 DM zur Verfügung. Wegen der räumlichen Trennung der Einrichtungen wird dieser Server für die konnektionistischen Verfahren nicht nutzbar sein. Er steht jedoch für die anderen Projektbereiche zur Verfügung. Weiterhin ist am Lehrstuhl projektrelevante Literatur im Wert von ca. 2000 DM vorhanden.

	2002	**2003**	**2004**
Für Sächliche Verwaltungsausgaben stehen als **Grundausstattung** voraussichtlich zur Verfügung:	75 000	75 000	75 000
Für Sächliche Verwaltungsausgaben werden als **Ergänzungsausstattung** beantragt (entspricht den Gesamtsummen “Sächliche Verwaltungsausgaben” in Übersicht 3.7):	20 700	4 500	4 500

(Alle Angaben in DM)

Begründung zur Ergänzungsausstattung der Sächlichen Verwaltungsausgaben

(515) *Kleingeräte:*

ungefährer Preis für einen PC Pentium Dual-III mit 1 GHz, 60 GB Festplatte und 1 GB Hauptspeicher. Der Rechner wird für die sehr zeit- und speicherintensiven Lernverfahren benötigt.

(522) *Verbrauchsmittel:*

jährliche Ausgaben:	3 000 DM	Büromaterialien, EDV-Zubehör und Software
	1 500 DM	ELRA-Mitgliedschaft
einmalige Ausgaben 2002:	4 800 DM	Softwarelizenzen für Lernverfahren
	5 400 DM	Lizenzen für digitale (Valenz-)Lexika (CELEX, IMSLex)

Reise-, Gast- und Vervielfältigungsmittel werden zentral beim Projekt Z beantragt.

3.7.3 Investitionen (Geräte über DM 20.000,- brutto und Fahrzeuge)

Es werden keine Investitionsmittel beantragt.