3.1. Allgemeine Angaben zum Teilprojekt A1

3.1.1. Thema:

Repräsentation und Erschließung linguistischer Daten

3.1.2. Fachgebiete und Arbeitsrichtung:

Computerlinguistik, Korpuslinguistik, regelbasierte, probabilistische und lernbasierte Annotationsverfahren, Syntax des Deutschen

3.1.3. Leiter:

Hinrichs, Erhard, Prof. Dr.	Kübler, Sandra, Dr.
geb. am 17.8.1954	geb. am 23.10.1968
Seminar für Sprachwissenschaft	Seminar für Sprachwissenschaft
Universität Tübingen	Universität Tübingen
Wilhelmstr. 19	Wilhelmstr. 19
72074 Tübingen	72074 Tübingen
Tel.: 07071 - 29 75446	Tel.: 07071 - 29 78490
Fax: 07071 - 29 5214	Fax: 07071 - 29 5214
E-Mail: eh@sfs.uni-tuebingen.de	E-Mail: kuebler@sfs.uni-tuebingen.de

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

[x] nein (Hinrichs) [x] ja, befristet bis zum 31.1.2009 (Kübler, wiss. Assistentin C1)

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
• Untersuchungen mit humanen embryonalen Stammzellen	[ ] 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 2001	277.9	13.9	0	291.8
2002	93.6	2.3	0	95.9
2003	97.2	2.3	0	99.5
2004	100.8	2.3	0	103.1
Zwischen-summe	569.5	20.8	0	590.3
2005	116.4	3.5	25	144.9
2006	116.4	5.5	0	121.9
2007	116.4	1.5	0	117.9
2008	116.4	1.5	0	117.9

(Beträge in Tausend EUR)

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 vorrangig flache syntaktische Strukturen (Chunks) und speziell der Verbalkomplex im Deutschen als eigenständiger Phänomenbereich behandelt. In der zweiten Projektphase (2002-2004) lag der Schwerpunkt auf der automatischen Annotation komplexer syntaktischer Strukturen, speziell von komplexen Phrasen, topologischen Feldern und grammatischen Funktionen. In beiden Projektphasen haben sich die eingesetzten Annotationsverfahren v.a. auf die Analyse von Simplexsätzen[1] beschränkt.

In der hier beantragten Projektphase (2005-2008) sollen die zentralen, offenen Fragen, die sich aus den bisherigen beiden Projektphasen ergeben, in empirischer und methodischer Hinsicht bearbeitet werden und das Projekt somit insgesamt zu einem Abschluss gebracht werden. Diese offenen Fragen konzentrieren sich in empirischer

[1]Der Begriff *Simplexsatz* entspricht der englischen Bezeichnung von *simplex clause* und soll Sätze bezeichen, die keine satzwertigen Einbettungen bzw. keine satzwertigen parataktischen Einzelglieder enthalten.

Hinsicht auf die Annotation komplexer Satzgefüge des Deutschen. Durch die Erkennung topologischer Felder, die für die Makrostruktur deutscher Sätze von zentraler Bedeutung sind, ist hierfür bereits eine wesentliche analytische Voraussetzung geschaffen worden.

Die Analyse komplexer Satzgefüge wird sich auf folgende zentrale Phänomenbereiche konzentrieren: die Erweiterung des bisherigen Inventars grammatischer Funktionen um Adjunkte, die Annotation von Parataxe, Hypotaxe und Koordinationsstrukturen, sowie die Auflösung anaphorischer Beziehungen auf Satz- und Diskursebene. Für die Annotation dieser Phänomenbereiche sollen die in den bisherigen Projektphasen entwickelten inkrementellen Verfahren weiterentwickelt werden. Ein potenzieller Nachteil derartiger inkrementeller Verfahren besteht darin, dass Fehler früherer Annotationsebenen sich in nachfolgenden Analyseschritten fortpflanzen. Dieses Defizit kann jedoch durch gezielte Fehlerkorrektur ausgeglichen werden, die bereits getroffene Annotationsentscheidungen durch zusätzliche Information in nachfolgenden Analyseebenen zu revidieren vermag.

In methodischer Hinsicht sind in den bisherigen Projektphasen inkrementelle und holistische Annotationsverfahren gleichrangig eingesetzt worden. Diese Nebenläufigkeit soll in der hier beantragten Projektphase fortgeführt werden, wobei die Bandbreite der verwendeten computerlinguistischen Annotationsverfahren, vor allem im Bereich des Maschinellen Lernens, signifikant erweitert werden. Ein besonderer Schwerpunkt soll dabei auf der Entwicklung hybrider und minimal überwachter Annotationsverfahren liegen.

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 des Deutschen, wie etwa die Erkennung grammatischer Funktionen und die Auflösung von Anaphora, theorieneutral zu erschließen?
2. Welche computerlinguistischen Analyseverfahren eignen sich für die korpuslinguistische Erschließung grammatischer Phänomene?
3. Welche Evaluationsverfahren 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 in der nächsten Projektphase geplante Annotation komplexer Sätze

zielführend sind. Entsprechend der sich daraus ergebenden Phänomenbereiche wird im Folgenden der Stand der Forschung zur Annotation von Parataxe und Hypotaxe (Abschnitt 3.3.2), von grammatischen Funktionen (Abschnitt 3.3.1), sowie zur Anaphernresolution (Abschnitt 3.3.3) beschrieben.

3.3.1. Grammatische Funktionen

Grammatische Funktionen wie *Subjekt* und *Objekt* sind theorieneutrale Begriffe, die zum Standardinventar empirischer Grammatikforschung gehören und gleichermaßen in fast jede syntaktische Theorie eingegangen sind. Worin sich verschiedene syntaktische Theorien unterscheiden, zeigt sich darin, welchen Status sie grammatischen Funktionen einräumen und welche Klassen von Funktionen sie unterscheiden.

Reis (1982) schlägt für das Deutsche vor, bei Komplementen nicht primär aufgrund von grammatischen Funktionen, sondern vielmehr nach ihrer Kasusmarkierung zu unterscheiden. Die am weitesten gefasste Charakterisierung von grammatischen Funktionen findet sich in verschiedenen Varianten der Dependenzgrammatik, die ein großes Inventar von Komplement und Adjunktfunktionen unterscheiden.

Im Projekt A1 sind grammatische Funktionen mit verschiedenen Methoden annotiert worden: mit memory-based learning Ansätzen ebenso wie mit regelbasierten Verfahren. Kübler (2002, 2004) verwendet einen holistischen Ansatz, bei dem die Konstituentenstruktur gleichzeitig mit den grammatischen Funktionen annotiert werden. Kouchnir (2004*a*) behandelt die Erkennung grammatischer Funktionen als eigenständiges Problem, das Teil einer inkrementellen Annotationsarchitektur ist. Dieser Ansatz überträgt den Ansatz von Buchholz (2002) für die Penn Treebank auf das Deutsche, wobei die Auswahl der relevanten Merkmale für die Klassifikation grammatischer Funktionen wesentlich modifiziert werden musste.

Die regelbasierten Verfahren (Müller, 2004*b*; Trushkina, 2004) dagegen konzentrieren sich auf die Erkennung von Komplementen auf der Grundlage von morphologischen Informationen, Valenzlexika und Stellungsregularitäten im Deutschen. Der Ansatz von Trushkina zielt auf eine dependenzgrammatische Analyse des Deutschen ab, wie sie für andere Sprachen mit ähnlichen computerlinguistischen Methoden von Aït-Mokhtar et al. (2002) für das Französische und von Oflazer (2003) für das Türkische angewandt wurden. Für das Deutsche finden sich weitere dependenzgrammatische Ansätze bei Schiehlen (2003) und bei Duchier (1999), jedoch mit jeweils anderer Zielsetzung. Schiehlen beschäftigt sich primär mit Fragen der semantischen Unterspezifizierung, während bei Duchier der Schwerpunkt vorrangig auf der Entwicklung einer Parsingarchitektur im Rahmen des Constraint Logic Programming liegt.

3.3.2. Parataxe, Hypotaxe und Koordinationsstrukturen

Die computerlinguistische Analyse komplexer Sätze in großen Textkorpora erfordert

die robuste Erkennung von Parataxe und Hypotaxe. Neben "klassischen" Haupt- und Nebensätzen mit ihren charakteristischen Stellungseigenschaften des finiten Verbs sind für das Deutsche als weiterer Typus sog. *uneingeleitete Nebensätze* im Sinne von Behaghel (1928) relevant, die V2-Stellung aufweisen, obwohl sie einem Hauptsatz untergeordnet sind. Letzerer Typus kommt häufiger in gesprochener als in geschriebener Sprache vor (vgl. Auer (1998)) vor. Entsprechende Belege finden sich jedoch auch in geschriebener Sprache, v.a. in Zeitungstexten wie dem taz Korpus, das die primäre Datengrundlage für das Projekt A1 bildet.

In der computerlinguistischen Forschung finden sich zwei Ansätze zur maschinellen Erkennung von Hypotaxe und Parataxe: inkrementelle Verfahren, die die Erkennung von Teilsätzen als eigenständigen Verarbeitungsschritt behandeln, sowie holistische Verfahren, bei denen die Teilsatzerkennung unter die Erkennung komplexer syntaktischer Strukturen subsumiert wird. Als typisches Beispiel für einen inkrementellen Ansatz ist der Shared Task *Clause Identification* (Tjong Kim Sang und Déjean, 2001) des *Fifth Workshop on Computational Language Learning* (CoNLL-2001) zu nennen. Zu den holistischen Verfahren sind v.a. PCFGs zu zählen, wie sie für das Deutsche von Dubey und Keller (2003) eingesetzt worden sind.

Eine wesentliche analytische Grundlage für die robuste Erkennung von Parataxe and Hypotaxe im Deutschen bildet die Erkennung topologischer Felder, die für die Makrostruktur deutscher Sätze von zentraler Bedeutung sind und die im Projekt A1 mit verschiedenen Methoden annotiert werden. Kübler (2004); Liepert (2003); Müller und Ule (2002); Veenstra et al. (2002) verfolgen einen inkrementellen Ansatz, während Ule (2003) einen PCFG-Ansatz verwendet, der bisher auf die Erkennung von topologischen Feldern optimiert ist.

Für das Deutsche wird ein probabilistischer Topologische-Felder-Parser außerdem von Frank et al. (2003) zur Vorstrukturierung der Eingabe eines mächtigeren HPSG-basierten Parsers genutzt. Sie erreichen damit einen beträchtlichen Geschwindigkeitszuwachs neben erhöhter Robustheit.

Alle o.g. inkrementellen Ansätze beschränken sich auf die reine Erkennung topologischer Felder, weisen den dadurch konstituierten Sätzen bzw. Teilsätzen jedoch keine syntaktischen Funktionen zu, d.h. sie unterscheiden z.B. nicht, ob es sich bei subordinierten Sätzen um Komplement- oder Adjunktsätze handelt. Diese Forschungslücke soll in der hier beantragten Projektphase von A1 geschlossen werden.

Neben Parataxe und Hypotaxe stellen Koordinationsstrukturen, besonders von komplexen Phrasen, von ungleichen Kategorien und in ellipitischen Strukturen einen wesentlichen, komplexitätsbildenden Faktor für die automatische Satzanalyse dar. Während die Erkennung parataktischer und hypotaktischer Satzbeziehungen in der Parsingliteratur einen breiten Raum einnimmt, finden sich zur computerlinguistischen Behandlung von Koordinationsstrukturen nur wenige Arbeiten, obwohl die Relevanz

und Schwierigkeit des Phänomensbereichs fast einmütig konstatiert wird. Robuste Parsingansätzen mit der Strategie, *islands of certainty* zu erkennen, beschränken sich zumeist auf die Erkennung einfacher Koordinationen, etwa von NPen (Müller, 2004*b*; Trushkina, 2004). Komplexere Koordinationsstrukturen werden in der computerlinguistischen Literatur v.a. unter dem Gesichtspunkt bestimmter Grammatikformalismen, z.B. in Tree Adjoining Grammars (Sarkar und Joshi, 1996) und in Combinatory Categorial Grammar (Hockenmaier und Steedman, 2002), behandelt. Werden robuste Verfahren überhaupt in den Blick genommen, wie von Clark und Curran (2003), so orientiert sich die Abdeckung primär an den für den verwendeten Grammatikformalismus theoretisch interessanten Konstruktionen wie etwa *non-constituent conjunction*.

Die für das Deutsche charakteristischen und theoretisch interessanten Varianten asymmetrischer Koordination (etwa im Falle von asymmetrischen F2-Koordination (Wunderlich, 1988; Höhle, 1990, 1991; Kathol, 1990) sowie sogenannte SLF-Koordination im Sinne von Höhle (1990)) sind jedoch als eigener Phänomenbereich unseres Wissens nach bislang keiner computerlinguistischen Analyse unterzogen worden. Ein weiterer aus computerlinguistischer Sicht weitgehend unerforschter Phänomenbereich betrifft die Interaktion von Koordination und Ellipsen.

3.3.3. Anaphernresolution

Anaphernresolution gehört zu den klassischen Forschungsgebieten in der Computerlinguistik. Während in der Allgemeinen Sprachwissenschaft Anaphern v.a. auf der Satzebene im Rahmen der Bindungstheorie untersucht worden sind, steht in computerlinguistischen Arbeiten die Diskursanaphora im Vordergrund. In empirischer Hinsicht beschränken sich die meisten Studien auf das Englische, v.a. auf Dialoge und hier besonders auf aufgabenorientierte Dialoge (engl. *task-oriented dialogues*) sowie auf Anaphern, deren Antezedenzien NPen sind. Eine Ausnahme stellt Byron (2002) dar, die auch Ereignisanaphora und Referenz auf abstrakte Diskursentitäten behandelt. Zur Anaphernresolution im Deutschen liegen nur wenige computerlinguistische Arbeiten vor, die v.a. von Strube et al. (Müller et al., 2002; Strube et al., 2002) vorgelegt worden sind. Neben Strube et al. (2002) ist der Ansatz von Kouchnir (2003, 2004*b*), der ein Boosting Verfahren verwendet, die einzige Studie zum Deutschen, die sich auf Lernverfahren stützt. Beide Studien gründen sich auf eine sehr eingeschränkte Textdomäne, wobei Kouchnir zeigt, dass das auf diesen Daten trainierte System sich nicht auf andere Textsorten übertragen lässt.

In methodischer Hinsicht gründen sich regelbasierte Ansätze zur Anaphernresolution v.a. auf Diskurstheorien, die die Interaktion von Diskursstrukturen und -segmenten und der Salienz von Diskursentitäten modellieren. Die *Centering Theory* (Grosz et al., 1995) nimmt hier eine herausragende Stellung ein, da sie als theoretische Grundlage für die Anaphernresolution in einer Reihe von Einzelsprachen verwendet wird, u.a. von Strube und Hahn (1999) für das Deutsche. Auf der Satzebene sind v.a. die Annah-

men der Bindungstheorie in computerlinguistische Anaphernmodellierung eingegangen (vg. etwa (Pinkal, 1991) und (Pollard und Sag, 1994)).

Beginnend mit McCarthy und Lehnert (1995) werden alternativ zu regelbasierten Ansätzen inzwischen auch Ansätze des Maschinellen Lernens zur Anaphernresolution eingesetzt. Neben Entscheidungsbäumen (McCarthy und Lehnert, 1995; Soon et al., 2001; Ng und Cardie, 2002; Strube et al., 2002; Yang et al., 2003) finden sich auch Ansätze des *maximum entropy*-basierten Lernens (Morton, 2000), des *memory-based* Lernens (Preiss, 2002), des Co-Trainings (Müller et al., 2002), sowie von probabilistischen Modellen (**?**). Die Merkmalsauswahl, auf die sich diese Studien gründen, beschränkt sich zumeist auf dieselbe Auswahl von syntaktischen Merkmalen, wie Kongruenz, grammatische Funktionen, textuelle Distanz zwischen Pronomen und Antezedens, und von semantischen Merkmalen, wie *named entity*, Belebtheit, Definitheit sowie semantische Klassenzugehörigkeit potenzieller Antezedenzien. Trotz ähnlicher Merkmalsauswahl ist ein direkter Vergleich der erzielten Ergebnisse fast unmöglich, da die Bandbreite der modellierten Anaphernphänomene von Studie zu Studie stark differiert. Hierauf hat zu Recht Byron (2001) hingewiesen.

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 die Zielsetzung des Projekts relevant und im Bewilligungszeitraum 2002-2004 entstanden sind.

Publikationen und Manuskripte

Hinrichs, E.W., S. Kübler, F.H. Müller und T. Ule (2002): „A Hybrid Architecture for Robust Parsing of German“, in *Proceedings of LREC 2002*, Las Palmas, Spanien.

Hinrichs, E.W. und J. Trushkina: (2004*a*): „Forging Agreement: Morphological Disambiguation of Noun Phrases“, *Journal of Language and Computation*. (im Erscheinen)

Hinrichs, E.W. und J. Trushkina (2004*b*): „Rule-based and Statistical Approaches to Morpho-syntactic Tagging of German“, in *Proceedings of Intelligent Information Systems 2004*, Zakopane, Polen.

Kouchnir, B. (2004*a*): *Knowledge-poor grammatical function assignment for German*, SfS, Universität Tübingen.

Kouchnir, B. (2004*b*): „A Machine Learning Approach for German Pronoun Resolution“, in *ACL/EACL 2004 Student Research Workshop*, Barcelona, Spanien.

Kübler, S. (2002): *Memory-Based Parsing of a German Corpus*, Dissertation, Universität Tübingen. Version vom 3. November 2002. Überarbeitete Version ange-

nommen in der Buchreihe “Natural Language Processing”, Amsterdam: John Benjamins.

Kübler, S. (2004): „Parsing Without Grammar–Using Complete Trees Instead“, *in* N. Nicolov, R. Mitkov, G. Angelova und K. Boncheva, (Hrsg.), *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*, Current Issues in Linguistic Theory, John Benjamins, Amsterdam. (im Erscheinen)

Liepert, M. (2003): „Topological Fields Chunking for German with SVM's: Optimizing SVM-parameters with GA's“, in *Proceedings of RANLP 2003*, Borovets, Bulgarien.

Müller, F. H. (2004*a*): „Annotating Grammatical Functions in German Using Finite-State Cascades“, in *Proceedings of COLING 2004*, Geneva, Switzerland.

Müller, F.H. (2004*b*): *A Finite State Approach to Shallow Parsing and Grammatical Functions Annotation of German*, Dissertation, Universität Tübingen. Abgabe bis 31.10.2004.

Müller, F.H. und T. Ule (2002): „Annotating topological fields and chunks – and revising POS tags at the same time“, *in Proceedings of COLING 2002*, Taipei, Taiwan, S. 695–701.

Telljohann, H., E.W. Hinrichs und S. Kübler (2003): *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*, SfS, Universität Tübingen.
URL: *http://www.sfs.uni-tuebingen.de/resources/sty.pdf*

Trushkina, J. (2004): *Morphological Disambiguation and Dependency Parsing for German*, Dissertation, Universität Tübingen. Abgabe bis 31.12.2004.

Trushkina, J. und E.W. Hinrichs (2004): „A Hybrid Model for Morpho-syntactic Annotation of German with a Large Tagset“, in *Proceedings of EMNLP 2004*, Barcelona, Spanien.

Ule, T. (2003): „Directed Treebank Refinement for PCFG Parsing“, in *Proceedings of TLT 2003*, Vaxjö, Schweden.

Ule, T. (2004): *Parsing syntaktischer Strukturen des Deutschen mit erweiterten PCFGs*, Dissertation, Universität Tübingen. Abgabe bis 31.12.2004.

Ule, T. und J. Veenstra (2004): „Making CFGs Probabilistic with Treebank Refinement“, in *Proceedings of CLIN 2003*, Antwerpen, Belgien.

Veenstra, J., F.H. Müller und T. Ule (2002): „Topological Fields Chunking for German“, *in Proceedings of CoNLL 2002*, Taipei, Taiwan, S. 56–62.

3.5. Arbeitsprogramm (Ziele, Methoden, Zeitplan)

3.5.1. Ziele

In der hier beantragten Projektphase (2005-2008) sollen komplexe Satzgefüge des Deutschen automatisch annotiert werden. Durch die Erkennung topologischer Felder, die für die Makrostruktur deutscher Sätze von zentraler Bedeutung sind, ist hierfür

bereits eine wesentliche analytische Voraussetzung geschaffen worden.

Die Analyse komplexer Satzgefüge wird sich auf die folgenden, inhaltlich eng aufeinander bezogenen Phänomenbereiche konzentrieren:

- die Erweiterung des bisherigen Inventars grammatischer Funktionen um Adjunkte
- die Annotation von Parataxe, Hypotaxe und Koordinationsstrukturen
- die Auflösung anaphorischer Beziehungen auf Satz- und Diskursebene

Die Erweiterung des Inventars der grammatischen Funktionen um Adjunkte ergibt sich unmittelbar aus der derzeitigen Projektphase (2002-2004), in der die Konzentration auf der Annotation von Komplementen lag. Um auch satzwertige Komplemente und Adjunkte erkennen zu können, müssen zusätzlich parataktische und hypotaktische Satzkonstruktionen annotiert werden. Eine adäquate Behandlung von Koordinationen stellt eine weitere notwendige Erweiterung der bisherigen Projektarbeit dar, da die Erkennung von Koordinationsstrukturen für die Analyse grammatischer Funktionen ebenso wie für die Analyse von Parataxe und Hypotaxe eine wichtige Grundlage bildet. Gleichzeitig ist die Annotation grammatischer Funktionen von zentraler Bedeutung für die Auflösung anaphorischer Beziehungen, da grammatische Funktionen ein wichtiges Merkmal bei der Auswahl potenzieller Antezedenzien darstellen.

Für die Annotation dieser Phänomenbereiche sollen die in den bisherigen Projektphasen entwickelten inkrementellen Verfahren weiterentwickelt werden. Ein potenzieller Nachteil derartiger inkrementeller Verfahren besteht darin, dass Fehler früherer Annotationsebenen sich in nachfolgenden Analyseschritten fortpflanzen. Dieses Defizit kann jedoch durch gezielte Fehlerkorrektur ausgeglichen werden, die bereits getroffene Annotationsentscheidungen durch zusätzliche Information in nachfolgenden Analyseebenen zu revidieren vermag.

In methodischer Hinsicht sind in den bisherigen Projektphasen inkrementelle und holistische Annotationsverfahren gleichrangig eingesetzt worden. Diese Nebenläufigkeit soll in der hier beantragten Projektphase fortgeführt werden, wobei die Bandbreite der verwendeten computerlinguistischen Annotationsverfahren, vor allem im Bereich des Maschinellen Lernens, signifikant erweitert werden. Ein besonderer Schwerpunkt soll dabei auf der Entwicklung hybrider und minimal überwachter Annotationsverfahren liegen. Für die drei zu untersuchenden Phänomenbereiche sollen folgende Verfahren eingesetzt werden:

- PCFGs kombiniert mit regelbasierter Extraktion flacher Baumstrukturen für minimal überwachtes Lernen.

- Chunkparsing kombiniert mit *k-nearest neighbor*-Lernen
- regelbasiertes Parsen von Dependenzstrukturen kombiniert mit statistischen Methoden und symbolischen Lernverfahren
- regelbasierte Constraint-Grammatik Verfahren kombiniert mit symbolischen Lernverfahren

Die ersten drei hybriden Verfahren sollen zur Annotation von grammatischen Funktionen, Koordinationen, Parataxe und Hypotaxe eingesetzt werden. Das letztgenannte Verfahren soll zur Anaphernresolution verwendet werden.

3.5.2. Methoden und Arbeitsprogramm

Die Phänomenbereiche

Grammatische Funktionen: In der gegenwärtigen Projektphase konzentriert sich die Annotation grammatischer Funktionen auf Simplexsätze und hier primär auf die Klasse der Komplemente. In der nächsten Projektphase soll der Schwerpunkt auf der Annotation von Adjunkten liegen und außerdem komplexe Satzgefüge mit einbezogen werden. Die Annotation von Adjunkten lässt sich nur zum Teil mit den gleichen grammatischen Merkmalen (v.a. Kasusinformation) und mit den gleichen linguistischen Wissenquellen (v.a. Valenzrahmen) bewerkstelligen, die für die Komplementerkennung erfolgreich eingesetzt wurden. Für die Erkennung von Adjunkten sind zusätzlich Informationen über lexikalische Köpfe und über Relationen zwischen lexikalischen Köpfen von primärer Bedeutung. Um die Anbindung von Adjunkten an nominale bzw. verbale Köpfe resolvieren zu können, sind Kopf-Kopf-Beziehungen zwischen Nomen/Verben mit nominalen Köpfen innerhalb der Adjunkt-PP zu berücksichtigen. Will man bei PP-Adjunkten zusätzlich zwischen verschiedenen Adjunktklassen (z.B. lokativen bzw. temporalen Modifikatoren) unterscheiden, sind Kopf-Kopf-Beziehungen zwischen Präposition und dem nominalen Kopf innerhalb der PP entscheidend.

Die genannten Kopf-Kopf-Beziehungen sollen aus großen, automatisch partiell annotierten Datenmengen extrahiert werden. Mit dem TüPP-D/Z-Korpus liegt bereits ein derartig annotiertes Korpus im Umfang von circa 11,5 Mio. Sätzen vor. Weitere Datenquellen können mit Annotationswerkzeugen aufbereitet werden, die in den ersten beiden Projektphasen bereits erstellt wurden. Um eine möglichst umfassende Abdeckung der genannten Kopf-Kopf-Beziehungen gewährleisten zu können, sollen zusätzlich nicht-überwachte Clusteringverfahren eingesetzt werden. Hierfür sollen lexikalische Konzepthierarchien aus GermaNet genutzt werden, um die Kohäsion von Klassen von Adjunktköpfen bzgl. des modifizierten Nomens oder Verbs zu ermitteln. Als Arbeitshypothese gilt dabei, dass die Kohäsion dann am größten ist, wenn ein

GermaNet Konzept mit hinreichender, aber nicht zu großer Abstraktionsstufe in einer Hyperonymkette als Filter verwendet wird.

Parataxe, Hypotaxe und Koordination: Eine der hauptsächlichen Fehlerquellen bei statistischen und lernbasierten Verfahren zur Annotation syntaktischer Strukturen liegt in der Analyse von Koordinationsstrukturen. Die Fehler werden durch unterschiedliche Auslöser verursacht: Fehler beim POS Tagging, falsche Entscheidungen bezüglich des Skopus einer Koordination oder falsche Entscheidungen bezüglich der Kategorie oder der grammatischen Funktion des Mutterknotens.

Taggingfehler treten vor allem bei mehrgliedrigen Konjunktionen wie z.B. bei *weder . . . noch* auf, da die zweiten Glieder oft ambig sind. Da der POS Tagger nur einen Kontext von 2–3 Wörtern betrachtet, ist die Entscheidung für ein Adverb oft wahrscheinlicher als für eine Konjunktion. Ein nicht lexikalisierter Parser hat dann keine Grundlage, um eine Koordination anzunehmen. Ein Beispiel für solch eine falsch erkannte Konjunktion findet sich in Abb. 1. Hier wurde *noch* vom POS Tagger als Adverb (ADV) annotiert, wodurch der Parser das zweite Konjunkt nicht als solches erkennen konnte und stattdessen eine Subordination postulierte.

Die Korrektur solcher Taggingfehler speziell bei mehrgliedrigen Konjunktionen ist daher ein wichtiger Schritt zur korrekten Analyse komplexer Koordinationen. Eine solche Korrektur kann vor dem Parsingschritt von einem spezialisierten Modul durchgeführt werden, das die endgültige Entscheidung bezüglich des POS Tags für mögliche mehrgliedrige Konjunktionen auf einen größeren Kontext bis hin zum kompletten Satz stützen kann. Im Beispiel in Abb. 1 erhöht das Vorkommen von *weder* die Wahrscheinlichkeit, dass *noch* das zweite Glied dieser Konjunktion bildet. Abb. 2 zeigt eine deutlich verbesserte, allerdings noch nicht vollständig korrekte Analyse.

Bei inkrementellen Parsingverfahren muss die Entscheidung über den Skopus einer Koordination meist schon relativ früh im Analyseprozess getroffen werden. In Abb. 2 z.B. hat der Parser fälschlicherweise die Nominalphrasen *Greie-Fuchs und Kühn* und *die passenden Sounds* koordiniert. Solche Fehlentscheidungen lassen sich nur verhindern, indem den Entscheidungen ein größerer Kontext zugrundegelegt wird. In dem vorliegenden Beispiel muss zum einen die mehrgliedrige Konjunktion *weder . . . noch* erkannt werden, zum anderen die Parallelität zwischen den Konjunkten. Eine Korrektur solcher Fehlertypen kann nur von einem Modul geleistet werden, das Zugriff auf den gesamten Satz ebenso wie auf die erste (oft inkorrekte) syntaktische Analyse hat. Zur Korrektur bietet sich hier eine Adaption der von Ruland (2000) eingeführten Baumtransformationen an. Solche Baumtransformationen sind ideal geeignet, im gesamten Baum Informationen über Parallelität etc. zu erkennen und die nötigen Änderungen in der Baumstruktur auszuführen. Jedoch gehen die nötigen Änderungen in Koordinationsstrukturen über das Transformationsinventar des ursprünglichen Ansatzes von Ruland hinaus, so dass dieses für die Bedürfnisse hier erweitert werden muss.

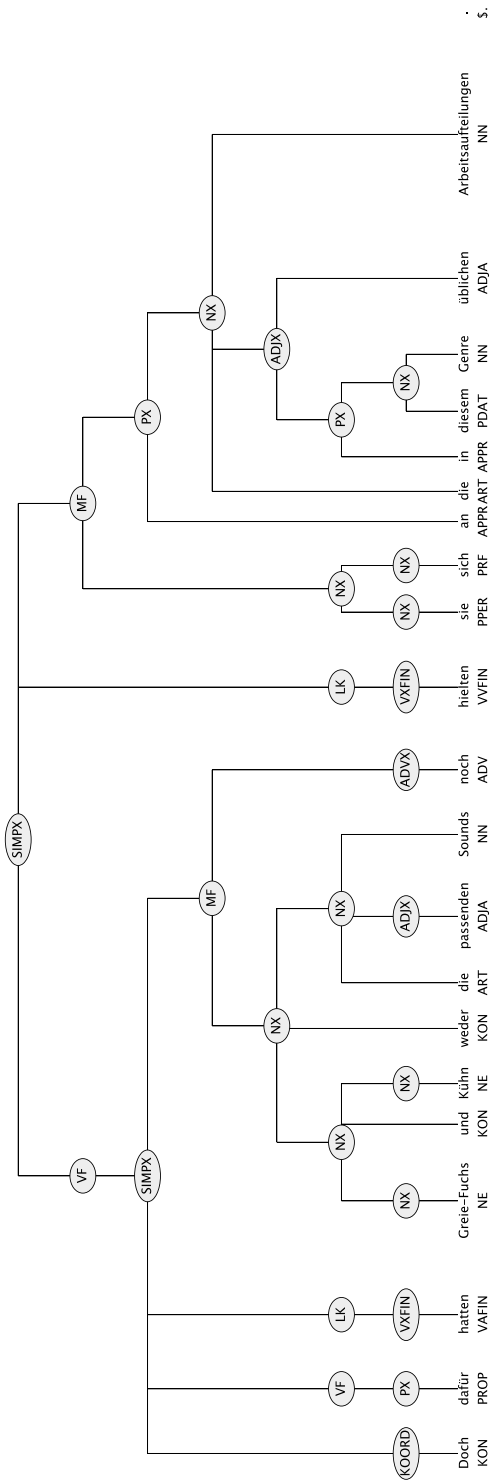

Abbildung 1: Eine falsch annotierte Koordination ausgelöst durch einen Taggingfehler.

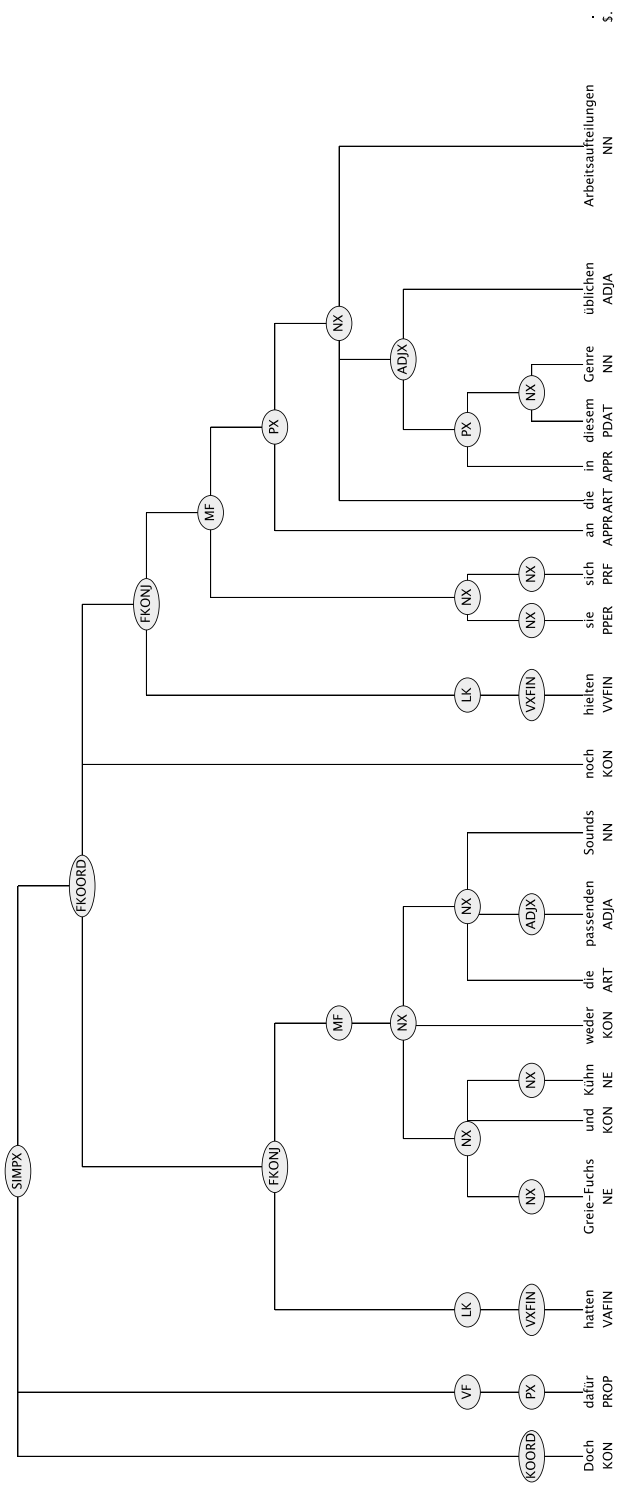

Abbildung 2: Die verbesserte Analyse mit korrektem POS Tag.

Koordinationsstrukturen finden sich auch auf der Ebene der Parataxe und lassen sich mit den gleichen Methoden behandeln, wie die oben beschriebenen Fälle von Felder- und Phrasenkoordinationen. Dies gilt auch für die nicht-koordinierenden parataktischen Strukturen mit dem Unterschied, dass in diesen Fällen automatisch annotierte Strukturen nicht so sehr transformiert, sondern durch parataktische Strukturen angereichert werden müssen.

Im Fall von hypotaktischen Strukturen sollen die grammatischen Funkionen der in Frage stehenden Teilsätze im Vordergrund stehen und hier speziell die Komplement-Adjunkt Unterscheidung. Da Elemente des C-Felds, wie subordinierende Konjunktionen und Relativpronomen, zu den häufigsten Fehlerquellen beim POS Tagging gehören, müssen diese zunächst mit Hilfe des topologischen Feldermodells einer systematischen Fehlerkorrektur unterzogen werden. Hinsichtlich der Komplement-Adjunkt Unterscheidung sollen die gleichen Verfahren eingesetzt werden, die für die Annotation nicht-satzwertiger grammatischer Funktionen erforderlich sind. Beim Clusteringverfahren muss für satzwertige grammatische Funktionen auf folgende Informationen zurückgegriffen werden: für Komplementsätze, wie z.B. *dass*-Sätze und freie Relativsätze, auf das C-Feld; für restriktive und nicht-restriktive Relativsätze auf den Kopf der modifizierten NP, die grammatische Funktion des Relativpronomens und das finite Verbs des Relativsatzes.

Anaphernresolution: Während sich die meisten Studien zur Anaphernresolution auf die Textsorte Dialog beschränkt, bildet das taz-Zeitungskorpus, das in den anderen beiden Pänomenbereichen des Projekts verwendet wird, die Datengrundlage auch für diesen Phänomenbereich. Diese Textsorte weist eine signifikant andere Vorkommensverteilung referenziell abhängiger Ausdrücke auf. Während in der gesprochenen Sprache Demonstrativa weit häufiger vorkommen als definite Beschreibungen, ist das Verhältnis bei Zeitungstexten genau umgekehrt. Daher soll der zu untersuchende Datenbereich neben der Auflösung von Antezedenzien für Personalpronomina, Reflexiva und Possessiva auch die referenzielle Abhängigkeit definiter Nominalphrasen und Demonstrativa einschließen. Neben den gängigen Relationen referenzieller Abhängigkeiten von Anaphora und Koreferenz soll außerdem ein weit feinkörnigeres Relationsinventar zugrundegelegt werden, das sich an den Empfehlungen des MATE Projekts zur Annotation von Diskursstrukturen (Davies et al., 1998) orientiert und das u.a. die Relationen *gebundene Variable, Teil-Ganzes-Beziehung* sowie *Element-Mengen-Relation* einschließt.

Will man ein derartig komplexes Daten- und Rolleninventar mit maschinellen Lernverfahren erfolgreich bearbeiten, so erfordert dies weitaus differenziertere linguistische Wissensquellen als jene Informationen, die in den meisten bisherigen Studien zur Anaphernresolution eingesetzt worden sind. Eine Ausnahme bilden hier **?**, die einen probabilistischen Parser verwenden, um Kopf-Konstituenten und deren semantische Typen zu extrahieren. Eine wesentliche Grundlage für derartig tiefes linguistisches

Wissen bilden die im Projekt A1 entwickelten automatischen Annotationsverfahren. Diese erlauben neben einer robusten Chunkanalyse die Zuweisung topologischer Felder und grammatischer Funktionen ebenso wie die morphologische Disambiguierung von Nominalphrasen. Für die Erkennung von *Named Entities* besteht mit der TüBa-D/Z bereits eine Datengrundlage zum überwachten Lernen, es müssen noch geeignete automatische Verfahren zu deren Analyse entwickelt werden.

Aufgrund der oben genannten Analyseverfahren lässt sich die Kandidatenmenge der Antezedenzien, die sog. *markables*, automatisch generieren und mit regelbasierten Verfahren filtern (siehe den Abschnitt zu den regelbasierten Constraint-Grammatik Verfahren kombiniert mit symbolischen Lernverfahren). Im Vergleich zu anderen Projekten, bei denen die markables manuell identifiziert werden müssen, lassen sich diese von der TüBa-D/Z Baumbank automatisch in ein zu erstellendes Korpus integrieren. Der verbleibende Annotierungsaufwand kann sich daher auf die Annotation der korrekten Antezedenzien und deren referenziellen Abhängigkeiten beschränken.

Neben der syntaktischen Struktur der markables ist insbesondere für die referenzielle Auflösung definiter NPen lexikalisch-semantische Information erforderlich. Bisherige Ansätze haben die dafür relevanten semantischen Eigenschaften auf wenige manuell annotierte Merkmale wie *belebt* und *abstrakt* reduziert. Ein derartiges Vorgehen reicht für die robuste Annotation definiter NPen in Zeitungskorpora jedoch nicht aus, wie das folgende Beispiel zeigt:

> Nach Angaben der Gewerkschaft IG Medien und des Verbandes **der Druckindustrie** erschienen der “Weser-Kurier” und die “Bremer Nachrichten” in eingeschränktem Umfang. Für die rund 200.000 Beschäftigten **der Branche** verlangt die IG Medien 6,5 Prozent mehr Lohn und Gehalt.

Um das Antezedens von “der Branche” korrekt identifizieren zu können, muss die Information verfügbar sein, dass es sich bei “Druckindustrie” um ein Hyponym von Branche handelt. Derartige Hyponomierelationen lassen sich aus Wortnetzen, wie das in Tübingen entwickelte GermaNet (Kunze und Wagner, 2001) für das Deutsche, extrahieren. Analoges gilt für *Teil-Ganzes-Beziehungen*, die in Wortnetzen als Meronymiebeziehungen modelliert werden.

Die hybriden Analyseverfahren

PCFGs kombiniert mit regelbasierter Extraktion flacher Baumstrukturen für minimal überwachtes Lernen: Das Problem zu spärlich vorhandener Daten ist allgegenwärtig in der linguistischen Verarbeitung, insbesondere bei der Behandlung komplexer syntaktischer Strukturen. Ausreichend große handannotierte Korpora sind nur mit sehr hohem Aufwand zu erstellen. Ein Ausweg aus diesem Dilemma ist die Kombination überwachter und unüberwachter Lernverfahren in einem hybriden Verfahren.

A1 Hinrichs/Kübler

Mit den beiden auf dem taz Zeitungskorpus basierenden Korpora TüBa-D/Z und TüPP-D/Z steht in Tübingen Datenmaterial zur Verfügung, das sehr gut für die Untersuchung hybrider Lernverfahren geeignet ist. Die TüBa-D/Z umfasst 266 441 Wörter in 15 260 Sätzen[2], und wurde manuell annotiert. Das Textmaterial wird in der TüBa-D/Z mit tief verzweigenden Bäumen auf Satzebene, der Ebene topologischer Felder sowie der Phrasenebene annotiert. Das TüPP-D/Z Korpus hingegen umfasst mit über 204 Mio. Wörtern in 11,5 Mio. Sätzen sehr viel mehr Daten, wurde aber automatisch annotiert. Hier kommt ein flaches, partielles Annotationsschema zum Einsatz. Die Verwendung des manuell annotierten TüBa-D/Z Korpus entspricht dabei überwachtem Lernen, die Verwendung von TüPP-D/Z unüberwachtem Lernen (da in keinem Schritt von der Datenaufbereitung bis zum Training manuelle Eingriffe vorgenommen wurden).

In der letzten Phase wurden unter anderem probabilistische kontextfreie Grammatiken (*Probabilistic Context-Free Grammars, PCFGs*) eingesetzt, um komplexe Phrasen und topologische Felder zu annotieren. In der kommenden Phase soll dieser Ansatz ausgeweitet werden auf die Annotation grammatischer Funktionen und komplexer syntaktischer Strukturen auf Satzebene. Grundlegend ist dabei die Frage nach der Auswahl der linguistischen Information, die nötig ist, um komplexe syntaktische Strukturen mit hoher Genauigkeit zu annotieren (siehe auch Hwa (1999)). Zur Annotation grammatischer Funktionen zum Beispiel ist sicherlich die Morphologie des Verbs von zentraler Bedeutung, da diese Rückschlüsse über die zu erwartende Argumentstruktur des Verbs und seiner Adjunkte erlaubt. Der Umfang der in der TüBa-D/Z manuell annotierten morphologischen Golddaten ist aber zu klein, um eine verlässliche (d.h. ausreichend große) Menge lernbarer morphologischer Trainingsinstanzen zu erhalten. Daher sollen unüberwachte Lernverfahren herangezogen werden, die die nötige Information aus dem TüPP-D/Z Korpus extrahieren können.

Als Konsequenz der verschiedenen Annotationsschemata in den beiden Korpora – tiefe Baumstruktur in der TüBa-D/Z gegenüber flachen partiellen Analysen in TüPP-D/Z – findet man deutliche Unterschiede in der Klammerung von Konstituenten und des Orts ihrer Anbindung am Baum. Dabei handelt es sich aber häufig um regelmäßige, vorhersagbare Unterschiede, die durch den bei der Annotation des TüPP-D/Z Korpus zum Einsatz gekommenen Formalismus entstanden sind. Es müssen also geeignete Abbildungen gefunden werden, die das Format der Daten in TüPP-D/Z in das Format der Datenstrukturen in TüBa-D/Z transformieren können, wobei aber nicht unbedingt alle Daten abgebildet werden müssen: Wie Pereira und Schabes (1992) beschreiben, kann auch nur die Verwendung partieller syntaktischer Analysen eine Leistungsverbesserung mit sich bringen. Es soll untersucht werden, welche Information, die in TüPP-D/Z enthalten ist, verwendet werden kann, um eine solche Leistungsverbesserung zu erzielen.

[2] In absehbarer Zeit sind Erweiterungen des Umfangs geplant.

Es ist hinreichend bekannt, dass die Mächtigkeit einfacher PCFGs nicht ausreicht, komplexe syntaktische Phänomene natürlicher Sprachen adäquat auszudrücken (insbesondere solche, deren korrekte Analyse die Einbeziehung von syntaktischem Kontext erfordert). Die Einbindung von mehr Kontext bringt eine deutliche Verbesserung der Leistung. Die Hinzunahme lexikalischer Information über Köpfe, d.h. die Lexikalisierung einer PCFG (*Head-Lexicalized Probabilistic Context-Free Grammars*, *HPCFGs*) erweitert signifikant die Vorhersagekraft einer PCFG, ohne den kontextfreien Formalismus vollständig zu verlassen. Deshalb wird die Lexikalisierung von PCFGs der erste Schritt hin zu einer Leistungsverbesserung sein. Letztlich muss aber auch die Mächtigkeit einer HPCFG durch ihre kontextfreie Basis beschränkt bleiben.

Es bietet sich an, diesen konzeptionsbedingten Nachteil durch geeignete Vor- bzw. Nachbearbeitungsschritte (d.h. vor bzw. nach der eigentlichen Anwendung eines PCFG Parsers bzw. Lerners) auszugleichen. Dies kann zum einen durch direkte Transformationen des Ausgangskorpus geschehen (Klein und Manning, 2001; Ule, 2003), indem Kontexteffekte, die durch PCFGs nicht erkannt werden können, schon von vornherein in Kategorien explizit dargestellt werden. Ebenfalls untersucht werden sollen zum anderen auch Verfahren der Nachbearbeitung bzw. Fehlerkorrektur. Ruland (2000) beschreibt ein Ersetzungssystem ähnlich dem Fehlerkorrekturverfahren des Brill Taggers, das falsche Teilbäume durch richtige Analysen ersetzt. Das System basiert auf einer Menge von Regeln, die aus dem Vergleich eines geparsten Satzes mit dem Goldstandard inferiert wurden.

Eine zentrale Voraussetzung für den Erfolg aller Ansätze ist die Verwendung und Verarbeitung sehr großer Datenmengen, die außerordentlich zeit- und ressourcenintensiv ist. Unabdingbar ist daher die Entwicklung paralleler Verarbeitungsalgorithmen, die solchen Ansprüchen gewachsen sind. Für deren Entwurf und Implementierung soll das beantragte hochperfomante Linux-Clustersystem eingesetzt werden.

Chunkparsing kombiniert *k-nearest neighbor*-Lernen: Das *k-nearest neighbor* (*k*-nn) Lernverfahren ist eine generellere Version des *memory-based learning*, bei dem die *k* ähnlichsten Instanzen zur Klassifizierung von neuen Instanzen herangezogen werden. Die Definition der Ähnlichkeitsfunktion unterliegt bei *k*-nn Verfahren weniger Beschränkungen als beim *memory-based learning*.

In der derzeitigen Phase (2002-2004) hat sich die Kombination von Chunkparsing und einem *memory-based* Lernansatz als besonders erfolgreich bezüglich der Annotationsgüte und der Robustheit erwiesen. Dieser Ansatz soll in der beantragten Phase auf eine breitere Basis gestellt werden, wobei die flache Chunk-Analyse weiterhin als Vorstrukturierung des Suchraums dient, die genügend Merkmale liefert, um eine Menge ähnlichster Bäume auszuwählen: Aus der Chunk-Analyse wird die Sequenz aus maximalen Chunks extrahiert. Diese Sequenz dient als Eingabe für die *k*-nn Suche nach dem ähnlichsten Syntaxbaum. Anders als beim bisherigen Ansatz wird die

lineare Abfolge der maximalen Chunks eine weniger exponierte Stellung einnehmen. Stattdessen wird das *Levenshtein-Distanzmaß*[3] (Kukich, 1992; Nerbonne et al., 1999) als Ähnlichkeitsfunktion eingesetzt. Die Levenshtein-Distanz berechnet die minimale Distanz zwischen zwei Strings bezüglich der Editierungsoperationen "Einsetzen" und "Löschen". Eine dritte, optionale Editierungsoperation ist das Ersetzen, das eine Löschaktion mit einer Einsetzaktion kombiniert. Dies ist bei Aufgaben sinnvoll, bei denen Ersetzen als weniger schwerwiegend betrachtet wird. Beim Einsatz der Levenshtein-Distanz zum Vergleich von Chunk-Sequenzen wird jede Kategorie eines maximalen Chunks als Zeichen des Strings betrachtet. Wegen der freien Stellung der Phrasen im deutschen Mittelfeld erscheint es sinnvoll, statt der Operation "Ersetzen" eine Operation "Vertauschen" zu definieren, die es erlaubt, dass zwei adjazente maximale Chunks ihre Position tauschen.

Jede Operation ist mit Kosten verbunden. Für die vorliegende Aufgabe müssen die optimalen Kosten für die Operationen "Einsetzen", "Löschen" und "Vertauschen" empirisch ermittelt werden. Dazu muß ein Maß gefunden werden, inwieweit sich ein Syntaxbaum verändert, wenn eine solche Operation durchgeführt wird. Als Ausgangspunkt bietet sich daher das in Kübler (2002) definierte Maß zur Baumähnlichkeit an.

Sind die *k* ähnlichsten Syntaxbäume durch ihre Ähnlichkeit in der Chunk-Analyse ermittelt, wird die Auswahl des ähnlichsten Baumes aufgrund lexikalischer Ähnlichkeiten ausgewählt. Die engste Definition von lexikalischer Ähnlichkeit besteht in der Identität der Headwörter der maximalen Chunks. Eine andere Möglichkeit besteht in der Ermittlung lexikalischer Clusters von Wörtern mit ähnlicher Distribution. Hier bietet sich der Einsatz von GermaNet (Kunze und Wagner, 2001) und das Verfahren zur Erkennung von Selektionspräferenzen von A. Wagner (Projekt C1) (Wagner, 2004) an.

Als letzter Schritt müssen die Levenshtein-Operationen, die zur Ermittlung des ähnlichsten Syntaxbaumes verwendet wurden, in Baummodifikationen umgesetzt werden, d.h. "Löschen" führt zum Löschen eines Teilbaumes, "Einsetzen" zum Einsetzen eines neuen Teilbaumes und "Vertauschen" zur strukturellen Veränderung des Baumes. Die Struktur des neuen Teilbaumes wird genauso ermittelt wie die interne Struktur der den maximalen Chunks entsprechenden Phrasen, falls diese unterschiedliche Strukturen aufweisen (vgl. (Kübler, 2002, Kap. 5.8)).

Beim Training dieses Verfahrens muss eine große Menge an Satzkombinationen verglichen werden. Da von den Einzelergebnissen dieser Vergleiche jedoch nur die *k* besten behalten werden müssen, ist es sinnvoll, dieses Verfahren mit der Entwicklung paralleler Analyseverfahren zu verbinden. Für diesen Trainingsalgorithmus soll das beantragte hochperformante Linux-Cluster eingesetzt werden.

[3] Dieses Maß ist auch unter dem dem Namen *Edit distance* bekannt.

Regelbasiertes Parsen von Dependenzstrukturen kombiniert mit statistischen Methoden und symbolischen Lernverfahren: In der letzten Phase wurde eine große Constraint-Grammatik für das regelbasierte Parsen von Dependenzstrukturen (Nivre, 2003) zur Verwendung mit dem *Xerox Incremental Parser (XIP)* (Aït-Mokhtar et al., 2002) entwickelt (Trushkina, 2004). In der kommenden Antragsphase soll dieser Ansatz kombiniert werden mit statistischen Methoden und symbolischen Lernverfahren. Dabei sollen diese Methoden sowohl in der Verarbeitung der zu parsenden Daten wie auch bei der Erstellung der Dependendenzgrammatik selbst zum Einsatz kommen.

Dependenzanalysen sind gegenüber klassischen baumorientierten Analysen besonders attraktiv, wenn komplexe Relationen zwischen Wörtern des analysierten Satzes repräsentiert werden sollen. So eignen sich Dependenzanalysen zum Beispiel besonders gut sowohl als Ausgangsbasis zur automatischen Annotation wie auch als Repräsentationsformat für Grammatische Funktionen, da hier die Relation eines Verbs zu seinen Komplementen und Adjunkten klar im Vordergrund steht. Ein Nachteil ist, dass die Grammatiken für das XIP-System vollständig manuell erstellt werden müssen. Dies bedeutet zwangsläufig, dass nur Phänomene aus beschränkten Datenmengen betrachtet werden können. Die Kombination von statistischen Verfahren und symbolischer Lernverfahren bieten auf zwei Ebenen einen Ausweg aus diesem Problem.

Statistische Verfahren sind sehr gut geeignet, Informationen aus großen Korpora für einen Dependenzparser nutzbar zu machen. In der letzten Phase wurden bereits erfolgreich statistische Ansätze zur morphologischen Desambiguierung mit einem nachgeschalteten XIP-Parser kombiniert. Diese Herangehensweise soll in der kommenden Antragsphase intensiviert betrachtet und auf die kombinierte Annotation komplexer Satzgefüge ausgeweitet werden.

Während die statistischen Verfahren auf den Ein- bzw. Ausgabedaten des inkrementellen Parsers operieren werden, sollen symbolische Lernverfahren eingesetzt werden, um die verwendeten Dependenzgrammatiken selbst zu lernen (Yamada und Matsumoto, 2003; Della Pietra et al., 1994; Nivre und Scholz, 2004). Der dem XIP-System zugrunde liegende Grammatikformalismus unterstützt eine solche Herangehensweise in besonderer Weise. Anders als in traditionellen Grammatikformalismen ist die Bedeutung einer einzelnen Regel nicht durch die Gesamtheit aller Regeln determiniert, sondern jede Regel bildet eine abgeschlosse, unabhängig interpretierbare Einheit (*self-containment*). Ob eine Regel angewandt werden kann, hängt ausschließlich davon ab, ob der in der Regel geforderte Kontext zum Betrachtungszeitpunkt erfüllt ist. Weiterhin erlaubt der XIP-Formalismus die Unterspezifikation von Informationen in Regeln. Dadurch können komplexe linguistische Phänomene auf mehrere Regeln verteilt werden, die nacheinander allgemeine und dann immer speziellere Fälle eines Phänomens beschreiben können (*descriptive decomposition*). Die Attraktivität dieser beiden Eigenschaften des XIP-Formalismus für symbolische Lernverfahren liegt darin, dass es möglich ist, für spezifische linguistische Probleme spezialisierte Lernverfahren einzu-

setzen, die jeweils klar umgrenzte Mengen von Regeln produzieren. Da jede Regel für sich angewandt wird, können alle gelernten Regelmengen zu einer großen Grammatik zusammen gefasst werden. Durch die Auftrennung in mehrere Teilprobleme wird die Komplexität der einzelnen Lernaufgabe wesentlich verkleinert.

Regelbasierte Constraint-Grammatik Verfahren kombiniert mit symbolischen Lernverfahren: Das im Folgenden beschriebene, hybride Verfahren soll für die Anaphernresolution angewendet werden. Ein wesentliches Problem für die Anwendung von Lernverfahren auf die Anaphernresolution besteht in der großen Menge negativer Trainingsinstanzen: dem jeweils korrekten Antezendens stehen eine Vielzahl potenzieller Kandidaten gegenüber. Dies hat zur Folge, dass Lernverfahren fälschlich übergeneralisieren und das korrekte Antezedens als weitere negative Instanz klassifizieren, da eine negative Entscheidung bei allen anderen Kandidaten die korrekte Entscheidung darstellt.

Die Performanz von Lernverfahren kann dadurch verbessern werden, dass die Anzahl der potenziellen Kandidaten durch ein vorgeschaltetes regelbasiertes Verfahren eingeschränkt wird. Für eine solche Filterung von Kandidaten bietet sich der Einsatz von Constraint-Grammatiken an, wie sie auch zur Analyse von Dependenzen (siehe vorhergehenden Abschnitt) in A1 eingesetzt werden. Aus diesem Grund soll, wie im Falle der Dependenzanalyse, das XIP System verwendet werden (vgl. (Trouilleux, 2001) für einen solchen Ansatz zum Französischen). Zur Filterung der Kandidatenmengen sollen die folgenden Arten von linguistischer Information verwendet werden: morphologische Information zur Numerus- und Genus-Kongruenz, syntaktische Strukturen und grammatische Funktionen. Diese Informationen stehen aufgrund eigener Vorarbeiten (Müller, 2004*b*; Trushkina, 2004; Ule, 2004) zur Verfügung und lassen sich unmittelbar in das XIP System integrieren.

Die tatsächliche Auswahl des wahrscheinlichsten Antezendens aus der gefilterten Liste erfolgt durch den Einsatz von Lernverfahren. Ebenso wie beim syntaktischen Parsing muss auch in diesem Problembereich überwachtes Lernen eingesetzt werden, wobei die Trainingsdaten im Projekt erstellt werden sollen. Aus dem Einsatz überwachter Lernverfahren ergibt sich auch bei der Anapherresolution das Problem zu geringer Trainingsdaten. Dieses Problem soll dadurch angegangen werden, dass zum einen Klassifikationsverfahren wie *memory-based learning* eingesetzt werden, für die empirisch nachgewiesen wurde, dass sie auf relativ kleinen Datenmengen gute Ergebnisse erzielen (vgl. (Banko und Brill, 2001; Kübler, 2004)). Zum anderen sollen diese Verfahren durch Ansätze zum minimal überwachten Lernen ergänzt werden. Hier bieten sich *ensemble learning* (Dietterich, 2002), und hier vor allem *boosting* (Schapire, 2002) an, mit dem Kouchnir (2004*b*) schon erfolgreich Personal- und Possessivpronomen im Deutschen resolvierte. Es soll jedoch auch ein *bootstrapping* Ansatz zur Vergrößerung der Menge von Trainigsdaten auf seine Verwendbarkeit für diese Aufgabe untersucht werden. Für die referenzielle Auflösung definiter NPen soll ebenso wie

für die Anbindungsdesambiguierung von PP-Adjunkten GermaNet eingesetzt werden, wobei für die Bestimmung der semantischen Ähnlichkeit zwischen Wortbedeutungen die spezifischen Eingenschaften von Wortnetzen berücksichtigt werden müssen (vgl. z.B. (Jiang und Conrad, 1997; Lin, 1998)).

3.5.3. Zeitplan

2005

- Grammatische Funktionen & Parataxe, Hypotaxe, Koordination:
  - Optimierung bestehender PCFG-Grammatiken auf Basis von TüBa-D/Z zur Behandlung von Komplementen und Adjunkten
  - Erstellung einer lexikalisierten PCFG auf Basis von TüBa-D/Z zur Behandlung von Komplementen und Adjunkten
  - Entwurf eines parallel arbeitenden PCFG-Parsingsystems zur Verarbeitung großer Korpora
  - Extraktion und Optimierung von für den *k*-nn Klassifikator relevanten Informationen aus der Chunk-Analyse
  - Implementierung des Standard *k*-nn Klassifikators
  - Erweiterung der XIP Dependenzgrammatik zur Erweiterung auf satzwertige Adjunkte
- Anaphernresolution:
  - Erstellung eines Moduls für *Named Entities*
  - Extraktion relevanter morphologischer und syntaktischer Strukturen aus der TüBa-D/Z als Grundlage für die Annotation und die regelbasierte Filterung
  - Annotation eines Korpus mit referenziellen Abhängigkeiten
  - Erstellung eines Stylebooks für die Korpusannotation
  - XIP-Implementierung der regelbasierten Filterung

2006

- Grammatische Funktionen & Parataxe, Hypotaxe, Koordination:
  - Auswahl und Bewertung von regelbasierten Verfahren und maschinellen Lernverfahren hinsichtlich ihrer Verwendbarkeit zur Leistungsverbesserung und Fehlerkorrektur von PCFG-Parses
  - Basisimplementierung des parallelen PCFG-Parsingsystems
  - Entwicklung geeigneter Ähnlichkeitsmaße zwischen Syntaxbäumen für das Training der Operationen der Levenshtein-Distanz im *k*-nn Klassifikator

  - Training der Gewichte für die Operationen der Levenshtein-Distanz
  - Optimierung der Baummodifikation im *k*-nn Klassifikator
  - Verwendung von TüPP-D/Z zur Auffindung von Wortpaaren zur Bestimmung von Präferenzen bei der PP-Anbindung
  - Verwendung der gewonnenen Wortpaare zur verbesserten automatischen PP-Annotation
  - Erweiterung der XIP Dependenzgrammatik auf nicht-satzwertige Adjunkte
  - Lernen von Dependenzgrammatiken auf der Basis von XIP-annotierten Daten und von TüBa-D/Z
- Anaphernresolution:
  - Fortführung der Korpusannotation
  - Implementierung eines hybriden Basissystems zur Resolution pronominaler Anaphern unter Verwendung von *memory-based learning* und *boosting*
  - Entwicklung geeigneter Ähnlichkeitsmaße zwischen GermaNet Konzepten

2007

- Grammatische Funktionen & Parataxe, Hypotaxe, Koordination:
  - Abschluss der Implementierung sowie Optimierung des parallelen PCFG-Parsingsystems; Kombination regelbasierter Verfahren mit dem Parser in ein hybrides Parsingsystem
  - Entwicklung einer parallelen Trainingskomponente für den *k*-nn Klassifikator
  - Ermittlung lexikalischer Cluster unter Einsatz von GermaNet für den *k*-nn Klassifikator und zur Auffindung von Hyperonymketten mit gemeinsamen Präferenzen bei der PP-Anbindung
  - Verwendung der gewonnenen Cluster zur verbesserten automatischen PP-Annotation; Integration beider Formalismen in die XIP und PCFG Parsingarchitekturen
  - Lernen von Dependenzgrammatiken auf der Basis von XIP-annotierten Daten und von TüBa-D/Z
- Anaphernresolution:
  - Fortführung der Korpusannotation unter Einbeziehung gesprochener Sprache (TüBa-D/S)
  - Unterstützung der manuellen Annotation durch die Implementierung eines *bootstrapping*-Verfahrens

  - Erweiterung des Basissystems um definite Beschreibungen und Demonstrative sowie im Bedarfsfall um weitere Lernverfahren

2008

- Grammatische Funktionen & Parataxe, Hypotaxe, Koordination:
  - Erweiterung des hybriden Parsingsystems um maschinelle Lernverfahren
  - Integration der PP-Cluster und der Wortpaare im *k*-nn KLassifikator
  - automatische Überführung der gelernten Dependenzen in XIP-Regeln
  - Anwendung der Parsingverfahren zur Annotation grammatischer Funktionen und komplexer Satzgefüge in großen Korpora
- Anaphernresolution:
  - Aktualisierung des Stylebooks hinsichtlich gesprochener Sprache
  - Abschluss der Annotationsarbeiten und Bereitstellung der Daten für externe Nutzer
  - Adaption des Systems für gesprochene Sprache
- gesamt:
  - Abschluss offener Arbeiten und Dokumentation

3.6. Stellung innerhalb des Sonderforschungsbereichs

- **A2 (Mönnich)**: Die Projekte A1 und A2 verbindet ein gemeinsames Interesse an den Konsequenzen, die die Wahl der linguistischen Datenstrukturen in Annotationssystemen hat. Während A1 diese Fragestellung primär aus der Sicht automatischer Annotationsverfahren betrachtet, liegt der Schwerpunkt in A2 auf der theoretischen Fundierung ausdrucksstarker Grammatikformalismen und Querysprachen für die Korpusrecherche.
- **A3 (Sternefeld)**: Der zentrale Forschungsgegenstand in A3 besteht in suboptimalen Strukturen und in der Gradienz von Grammatikalität, denen als Datentyp intuitive Sprecherurteile zugrundeliegen. Eben solche suboptimalen Strukturen finden sich auch in dem von A1 verwendeten Zeitungskorpus der taz, das häufig Charakteristika der Spontansprache aufweist. Dadurch ergibt sich die Chance eines Datentypenvergleichs von intuitiven Sprecherurteilen und Korpusdaten bezüglich der Gradienz von Grammatikalität.
- **A4 (Pafel)**: Bei den Cross-modal Priming Experimenten in A4 nimmt die Anaphernresolution und die Aktivierung von Diskursreferenten eine Schlüsselrolle ein. Diese Fragestellungen haben ihr Gegenstück in A1 in der automatischen Anaphernresolution, bei denen die diskursstrukturelle Salienz potenzieller Antezedenzien eine zentrale Bedeutung hat. Wie im Falle von A3 ergibt

sich hieraus die Chance eines Datentypenvergleichs von experimentellen und Korpusdaten in Bezug auf den gemeinsamen Phänomenbereich der Anaphern. Ferner wird A1 das Projekt A4 durch die Bereitstellung annotierter Daten bei der Korpusrecherche zu Quantorenskopuslesarten unterstützen.

- **A5 (Richter)**: Die Projekte A1 und A5 verbindet ein gemeinsames Interesse an partiell annotierten Korpusdaten, phänomenorientierter Grammatikentwicklung und an Constraintgrammatiken. Während letztere in A1 eher dependenzgrammatisch fundiert ist, wird in A5 mit der HPSG v.a. eine phrasenstrukturbasierte Theorie verfolgt.
- **B3 (Ehrich/Reich)**, **B13 (Winkler)** und **B15 (Reis/Truckenbrodt)**: Diese Teilprojekte haben ihre empirischen Schwerpunkte auf der Analyse komplexer Satzgefüge (B15) von Koordinationsstrukturen (B3,B13) und von Ellipsen (B13), die sich mit den von A1 bearbeiteten Phänomenbereichen jeweils signifikant überlappen, wobei B3 und B15 ebenfalls hauptsächlich Daten des Deutschen untersuchen. Zusätzlich verbindet A1 mit B3 ein Interesse an Verarbeitungsstrategien, wobei der Begriff paralleler Strukturen als strukturgebendes Element in beiden Projekten eine wesentliche analytische Grundlage bildet. Im Rahmen dieser Kooperation wird sich A1 auch an den geplanten Workshops zu *Koordination* und zu *Complex Clauses - Linguistic, Psycholinguistic, and Computational Perspectives* (siehe Projekt Z) beteiligen.
- **B6 (Koch)**: Während lexikalisch-kognitive Relationen in B6 primär unter dem Motivationsaspekt betrachtet werden, spielen sie bei A1 für die bei der Analyse grammatischer Funktionen notwendigen Clusterverfahren ebenfalls eine entscheidende Rolle. Es besteht daher ein komplementäres Interesse an Meronymie-, Pertonymie- bzw. Hyperonymie-Relationen, die für die Modellierung lexikalischer Relationen von Wortbedeutungen in Wortnetzen, wie WordNet and GermaNet, von grundlegender Bedeutung sind.
- **B11 (Butzenberger)**: In diesem Projekt wird ein Korpus des Tibetischen annotiert und u.a. hinsichtlich anaphorischer Beziehungen und satzübergreifender Relationen systematisch erweitert. Hierbei werden sich sprachübergreifende Gesichtspunkte hinsichtlich der verwendeten linguistischen Kategorien und der von A1 entwickelten Annotationsebenen ergeben. Andererseits sind durch die verwandten Phänomenbereiche von B11 wichtige Impulse für die von A1 zu entwickelnden linguistischen Repräsentationen zu erwarten.
- **C1 (Reis/Hinrichs)**: Fortsetzung und Abschluss der Annotationsarbeiten von A1 soll in enger Kooperation mit C1 erfolgen, wobei die laufende Integration der annotierten Daten in das TUSNELDA Format im Vordergrund stehen wird. In Kooperation mit C1 stellt das Projekt A1 den anderen Einzelprojekten weiterhin Expertise bei der Enkodierung von Textkorpora in Standardformaten, speziell in den Auszeichnungssprachen XML und XHTML, zur Verfügung.

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:

- Center for Dutch Language and Speech, Antwerpen, und Induction of Linguistic Knowledge am Center for Language Studies, Tilburg: Der bereits bestehende Kontakt zu Prof. Dr. Walter Daelemans auf dem Gebiet des *memory-based learning* soll in der nächsten Phase fortgesetzt werden.
- Xerox Research Center Europe (XRCE), Grenoble: Die Kooperation mit Jean-Pierre Chanod, Salah Aït-Mokthar und Claude Roux zur Adaption des *Xerox Incremental Parsers* (XIP) für das Deutsche soll fortgesetzt werden.
- European Media Laboratory (EML) Research gGmbH, Heidelberg: Es wird eine Kooperation mit Dr. Michael Strube, dem Leiter der NLP-Gruppe am EML Research, zur Annotation und Auflösung von Anaphern, vor allem im Deutschen, angestrebt.
- School of Humanities, Languages and Social Studies, University of Wolverhampton: Die Kooperation mit Prof. Dr. Ruslan Mitkov und seiner Forschergruppe bezüglich regelbasierter Verfahren zur Anaphernresolution wird fortgeführt.
- Department of Computer Science, Ohio State University, Columbus: Es wird eine Kooperation mit Prof. Dr. Donna Byron zur Spezifikation von Anaphoraklassen und deren Annotation, sowie der Entwicklung von Evaluationsverfahren für Anaphernresolution angestrebt.

3.7. Abgrenzung gegenüber anderen geförderten Projekten

Kompetenzzentrum für Text- und Informationstechnologie (BMWK, 1.10.2000 - 30.9.2005):

Das in Kooperation mit der Universität Stuttgart unterhaltene Kompetenzzentrum beschäftigt sich mit der Entwickung und Pflege sprechtechnologischer Werkzeuge und Ressourcen. Die dort erstellten Ressourcen TüPP-D/Z und GermaNet stehen als Wissensquellen für die hier beantragte Phase des SFB 441 zur Verfügung. Die manuell annotierte Baumbank TüBa-D/Z wurde ebenfalls im Kompetenzzentrum erstellt und in Kooperation mit dem SFB 441 hinsichtlich SFB-spezifischer Bedürfnisse (Erweiterung durch morphologische Merkmale und Einbindung in TUSNELDA) erweitert. Nach Beendigung des Kompetenzzentrums müssen alle für die hier beantragte Projektphase nötigen Erweiterungen aus Ressourcen des SFB 441 bestritten werden.

Zitierte Literatur

Aït-Mokhtar, S., J.-P. Chanod und C. Roux (2002): „Robustness beyond shallowness: incremental deep parsing“, *Natural Language Engineering* 8(2–3), 121–144.

Auer, P. (1998): *Zwischen Parataxe und Hypotaxe: 'abhängige Hauptsätze' im gesprochenen und geschriebenen Deutsch*, Interaction and Linguistic Structures 2, Universität Konstanz.

Banko, M. und E. Brill (2001): „Scaling to Very Very Large Corpora for Natural Language Disambiguation“, in *Proceedings of ACL/EACL 2001*, Toulouse, Frankreich, S. 26–33.

Behaghel, O. (1928): *Deutsche Syntax*, Winter Verlag, Heidelberg.

Buchholz, S. (2002): *Memory-Based Grammatical Relation Finding*, Dissertation, Tilburg University.

Byron, D. (2001): „The Uncommon Denominator“, *Computational Linguistics* 27(4).

Byron, D. (2002): „Resolving Pronominal Reference to Abstract Entities“, in *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

Clark, S. und J. R. Curran (2003): „Log-Linear Models for Wide-Coverage CCG Parsing“, in *Proceedings of the SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP '03)*, Sapporo, Japan, S. 97–104.

Davies, S., M. Poesio, F. Brunesseaux und L. Romary (1998): *Annotating Coreference in Dialogues: Proposal for a Scheme for MATE*, MATE.
URL: *http://www.hcrc.ed.ac.uk/∼poesio/MATE/anno_manual.html*

Della Pietra, S., V. Della Pietra, J. Gillett, J. Lafferty, H. Printz und L. Ureš (1994): „Inference and Estimation of a Long-Range Trigram Model“, in *Proceedings of the Second International Colloquium on Grammatical Inference and Applications*, Band 862 von *Lecture Notes in Artificial Intelligence*, Springer, S. 78–92.

Dietterich, T. G. (2002): „Ensemble Learning“, *in* M. Arbib (Hrsg.), *The Handbook of Brain Theory and Neural Networks*, Second Edition Auflage, MIT Press, Cambridge, MA, S. 405–406.

Dubey, A. und F. Keller (2003): „Probabilistic Parsing for German using Sister-Head Dependencies“, in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, S. 96–103.

Duchier, D. (1999): „Axiomatizing Dependency Parsing Using Set Constraints“, in *Proceedings of the Sixth Meeting on Mathematics of Language (MOL 6)*, Orlando, FL, S. 115–126.

Frank, A., M. Becker, B. Crysmann, B. Kiefer und U. Schäfer (2003): „Integrated Shallow and Deep Parsing: TopP Meets HPSG“, in *Proceedings of ACL 2003*, Sap-

poro, Japan.

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

Grosz, B., A. Joshi und S. Weinstein (1995): „Centering: A Framework for Modelling the Local Coherence of Discourse“, *Computational Linguistics* 2(21).

Hockenmaier, J. und M. Steedman (2002): „Generative Models for Statistical Parsing with Combinatory Categorial Grammar“, in *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

Höhle, T. (1991): „On Reconstruction and Coordination“, *in* H. Haider und K. Netter (Hrsg.), *Representation and Derivation in the Theory of Grammar*, Band 22 von *Studies in Natural Language and Linguistic Theory*, Kluwer, Dordrecht, S. 139–197.

Höhle, T. N. (1990): „Assumptions about asymmetric coordination in German“, *in* J. Mascaró und M. Nespor (Hrsg.), *Grammar in progress. Glow essays for Henk van Riemsdijk*, Foris, Dordrecht, S. 221–235.

Hwa, R. (1999): „Supervised Grammar Induction using Training Data with Limited Constituent Information“, in *Proceedings of the 37th Annual Meeting of the ACL*, S. 73–79.

Jiang, J. J. und D. W. Conrad (1997): „Semantic Similarities Based on Corpus Statistics and Lexical Taxonomy“, in *Proceedings of the International Conference on Research in Computational LInguistics (ROCLING X)*, Taiwan.

Kathol, A. (1990): „Linearization vs. phrase structure in German coordination constructions“, *Cognitive Linguistics* 10(4), 303–342.

Klein, D. und C. D. Manning (2001): „Parsing with Treebank Grammars: Empirical Bounds, Theoretical Models, and the Structure of the Penn Treebank“, in *Proceedings of the 39th Annual Meeting of the ACL*.

Kouchnir, B. (2003): *A Machine Learning Approach to German Pronoun Resolution*, Magisterarbeit, School of Informatics, University of Edinburgh.

Kukich, K. (1992): „Techniques for Automatically Correcting Words in Text“, *ACM Computing Surveys* 24(4), 377–439.

Kunze, C. und A. Wagner (2001): „Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche“, *in* I. Lemberg, B. Schröder und A. Storrer (Hrsg.), *Chancen und Perspektiven computergestützter Lexikographie*, Band 107 von *Lexicographica Series Major*, Niemeyer, Tübingen, S. 229–246.

Lin, D. (1998): „An Information-Theoretic Definition of Similarity“, in *Proceedings of ICML 1998*, Madison, Wisconsin.

McCarthy, J. F. und W. G. Lehnert (1995): „Using decision trees for coreference re-

solution“, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI’95)*, Montreal, Canada, S. 1050–1055.

Morton, T. S. (2000): „Coreference for NLP Applications“, in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL’00)*, Hong Kong.

Müller, C., S. Rapp und M. Strube (2002): „Applying co-training to reference resolution“, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL’02)*, Philadelphia, PA, USA, S. 352–359.

Nerbonne, J., W. Heeringa und P. Kleiweg (1999): „Edit Distance and Dialect Proximity“, *in* D. Sankoff und J. Kruskal (Hrsg.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.*, CSLI, Stanford, CA, S. v–xv.

Ng, V. und C. Cardie (2002): „Improving machine learning approaches to coreference resolution“, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL’02)*, Philadelphia, PA, USA, S. 104–111.

Nivre, J. (2003): „An Efficient Algorithm for Projective Dependency Parsing“, in *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)*, Nancy, S. 149–160.

Nivre, J. und M. Scholz (2004): „Deterministic Dependency Parsing of English Text“, in *Proceedings of COLING 2004*, Genf.

Oflazer, K. (2003): „Dependency Parsing with an Extended Finite-State Approach“, *Computational Linguistics* 29(4), 515–544.

Pereira, F. und Y. Schabes (1992): „Inside-Outside Reestimation from Partially Bracketed Corpora“, in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware, S. 128–135.

Pinkal, M. (1991): „On the syntactic-semantic analysis of bound anaphora“, in *Proceedings of EACL 1991*, Berlin, S. 45–50.

Pollard, C. und I. Sag (1994): *Head-Driven Phrase Structure Grammar*, University of Chicago Press, Chicago, IL.

Preiss, J. (2002): „Anaphora resolution with memory based learning“, in *Proceedings of the 5th UK Special Interest Group for Computational Linguistics (CLUK5)*, S. 1–8.

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

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

Sarkar, A. und A. Joshi (1996): „Coordination in Tree Adjoining Grammars: Formalization and Implementation“, in *Proceedings of the 16th International Conference on Computational Linguistics: COLING 1996*, S. 610–615.

Schapire, R. E. (2002): „The boosting approach to machine learning: an overview“, in *Proceedings of the MSRI Workshop on Nonlinear Estimation and Classification*.

Schiehlen, M. (2003): „Combining Deep and Shallow Approaches in Parsing German“, in *Proceedings of ACL 2003*, Sapporo, Japan.

Soon, W. M., H. T. Ng und D. C. Y. Lim (2001): „A machine learning approach to coreference resolution of noun phrases“, *Computational Linguistics* 27(4), 521–544.

Strube, M., S. Rapp und C. Müller (2002): „The influence of minimum edit distance on reference resolution“, in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, Philadelphia, PA, USA, S. 312–319.

Strube, M. und U. Hahn (1999): „Functional centering: grounding referential coherence in information structure“, *Computational Linguistics* 25(3), 309–344.

Tjong Kim Sang, E. F. und H. Déjean (2001): „Introduction to the CoNLL-2001 Shared Task: Clause Identification“, *in* W. Daelemans und R. Zajac (Hrsg.), *Proceedings of CoNLL-2001*, Toulouse, France, S. 53–57.

Trouilleux, F. (2001): *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*, Dissertation, Université Blaise-Pascal, Clermont-Ferrand.

Wagner, A. (2004): *Learning Thematic Role Relations for Lexical Semantic Nets*, Dissertation, Universität Tübingen. Abgabe bis 10.11.2004.

Wunderlich, D. (1988): „Some Problems of Coordination in German“, *in* U. Reyle und C. Rohrer (Hrsg.), *Natural Language Parsing and Linguistic Theories*, Studies in Linguistics and Philosophy, Reidel, Dordrecht, S. 289–316.

Yamada, H. und Y. Matsumoto (2003): „Statistical Dependency Analysis with Support Vector Machines“, in *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)*, Nancy, S. 195–206.

Yang, X., G. Zhou, J. Su und C. Tan (2003): „Coreference Resolution Using Competition Learning Approach“, in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, Sapporo, Japan, S. 176–183.

Es bedeuten:
PK: Personalbedarf und –kosten (Begründung vgl. 3.8.1)
SV: Sächliche Verwaltungsausgaben (Begründung vgl. 3.8.2)
I: Investitionen (Geräte über EUR 10.000,– brutto; Begründung vgl. 3.8.3)

	Bewilligung 2004			2005			2006			2007			2008		
PK	Verg.-Gr.	An-zahl	Betrag in EUR	Verg.-Gr.	An-zahl	Betrag in EUR	Verg.-Gr.	An-zahl	Betrag in EUR	Verg.-Gr.	An-zahl	Betrag in EUR	Verg.-Gr.	An-zahl	Betrag in EUR
	BAT IIa	0	0	BAT IIa	1	58.800	BAT IIa	1	58.800	BAT IIa	1	58.800	BAT IIa	1	58.800
	BAT IIa/2	3	82.800	BAT IIa/2	1	27.600	BAT IIa/2	1	27.600	BAT IIa/2	1	27.600	BAT IIa/2	1	27.600
	Wiss. HK	1	18.000	Wiss. HK	1	18.000	Wiss. HK	1	18.000	Wiss. HK	1	18.000	Wiss. HK	1	18.000
	Stud. HK	0	0	Stud. HK	1	12.000	Stud. HK	1	12.000	Stud. HK	1	12.000	Stud. HK	1	12.000
	zusammen		100.800	zusammen		116.400	zusammen		116.400	zusammen		116.400	zusammen		116.400
SV				Kostenkategorie oder Kennziffer		Betrag in EUR	Kostenkategorie oder Kennziffer		Betrag in EUR	Kostenkategorie oder Kennziffer		Betrag in EUR	Kostenkategorie oder Kennziffer		Betrag in EUR
				515		(5.000)†	515		0	515		0	515		0
				522		1.500	522		1.500	522		1.500	522		1.500
				547		2.000	547		4.000	547		0	547		0
				zusammen		3.500	zusammen		5.500	zusammen		1.500	zusammen		1.500
I				Mittel für Invest. insgesamt:			Mittel für Invest. insgesamt:			Mittel für Invest. insgesamt:			Mittel für Invest. insgesamt:		
						25.000			0			0			0

†Zu beschaffen aus der Grundausstattung.

3.8.1. 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						
3.8.1.1	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, Dr.	Comp.-Ling. ML	SfS	5		
(einschl. HK)	4. Saile, Jochen, M.A.	Systemprog.	SfS	4		
3.8.1.2 nichtwiss. Mitarbeiter	Stoll, Cornelia	Sekretariat	SfS	5		
Ergänzungsausstattung						
3.8.1.3	5. N.N.	Masch. Lernen		41		BAT IIa
wissenschaftl.	6. Holger Wunsch, M.A.	PCFGs		20,5		BAT IIa/2
Mitarbeiter	7. Eva Klett	XIP		20		wiss. HK
(einschl. HK)	X 8. N.N.	Comp.-Ling.		20		stud. HK
3.8.1.4 nichtwiss. Mitarbeiter						

(Stellen, für die die Mittel <u>neu</u> beantragt werden, sind mit **X** gekennzeichnet)

A1 Hinrichs/Kübler

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, Lehrtätigkeit, Mitbetreuung von Abschlussarbeiten.

3. Leitung des Teilprojekts und Betreuung der damit verbundenen Forschungsarbeiten; Lehrtätigkeit, Mitbetreuung von Abschlussarbeiten.

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” mitarbeitet. Hierfür sind Erfahrung im Umgang mit Korpora, Taggern und maschinellen Lernverfahren notwendig.

   Diese Stelle soll mit einer/m promovierten Computerlinguistin/Computerlinguisten besetzt werden. Voraussetzung für diese Stelle sind fundierte Kenntnisse symbolischer Lernverfahren (speziell Memory-based Learning, k-nearest Neighbors) und Erfahrung mit computerlinguistischen Verfahren zur Anaphernresolution. Aufgrund der Vielschichtigkeit der vorausgesetzten Kenntnisse ist eine abgeschlossene Promotion erforderlich.

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

   Es ist vorgesehen, diese Stelle mit Herrn Holger Wunsch zu besetzen. Herr Wunsch besitzt ein fundiertes Wissen über den Einsatz statistischer Verfahren in der Computerlinguistik, insbesondere bei der Implementierung von PCFGs. Sein Dissertationsvorhaben “Minimal überwachte Optimierung von PCFGs mittels einer massiv parallelen Parsingarchitektur” steht in unmittelbarem Zusammenhang mit den geplanten Projektarbeiten. Seine Pilotstudie zum minimal überwachten Lernen von PCFGs bildet Grundlage und Motivation für die geplante massiv parallele Parsingarchitektur auf dem beantragten Linux-Cluster.

7. Zu besetzen durch eine(n) Computerlinguistikstudentin/en mit abgeschlossenem Studium

   Aufgabenbereich: Implementierung einer Dependenzgrammatik des Deutschen unter Verwendung der Xerox Parsing Platform XIP.

   Voraussetzung: Für diese Aufgabe ist Erfahrung auf dem Gebiet der Implementierung großer Grammatiken nötig. Es werden Programmierkenntnisse in Java und fundierte Kenntnisse zur Syntax des Deutschen vorausgesetzt.

   Es ist vorgesehen, diese Stelle mit Frau Eva Klett zu besetzen. Frau Klett verfügt bereits über einschlägige Projekterfahrung mit der Annotation morphologischer Merkmale in der TüBa-D/Z.

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

   Aufgabenbereich: Aufbau einer Sprachressource für die Anaphernresolution, in der Anaphern-Antezedens Beziehungen annotiert sind, unter Verwendung des am EML entwickelten Annotationswerkzeugs MMAX.

   Voraussetzung: Für diese Aufgabe sind fundierte Kenntnisse zur Syntax and Semantik des Deutschen erforderlich.

3.8.2. Aufgliederung und Begründung der Sächlichen Verwaltungsausgaben (nach Haushaltsjahren)

An Kleingeräten stehen den Projektleitern zwei Rechnerarbeitsplätze (IBM Laptops mit Flatscreens) und ein Netzwerkdrucker im Wert von insgesamt EUR 11 500 zur Verfügung.

Das Projekt kann außerdem, wie auch schon in den bisherigen Phasen, auf die am Lehrstuhl Allgemeine Sprachwissenschaft/Computerlinguistik vorhandene Rechnerausstattung und auf lizenzensierte Software für nicht-parallele symbolische Lernverfahren zurückgreifen. Hierfür stehen ein Dual Pentium III und ein Pentium IV Hyperthreading Rechner im Wert von EUR 7 000 zur Verfügung. Das SfS verfügt außerdem über zwei Dual-Opteron Server im Wert von insgesamt EUR 14 000, die in das beantragte Linux-Cluster (siehe Abschnitt 3.8.3) integriert werden sollen. Weiterhin ist am Lehrstuhl projektrelevante Literatur im Wert von ca. EUR 5 000 vorhanden.

Die Rechnerarbeitsplätze (mit Pentium II Prozessor und 333 bzw. 400 MHz) der Projektmitarbeiter wurden zu Beginn der ersten Projektphase angeschafft. Sie entsprechen inzwischen nicht mehr den gegenwärtigen technischen Anforderungen hinsichtlich Rechnerleistung, Speicher- und Plattenbedarf. Es werden daher im Haushaltsjahr 2005 vier Arbeitsplätze und zwei Monitore im Gesamtwert von EUR 5 000 (für zwei Projektmitarbeiterstellen und zwei Hilfskraftstellen (geprüft und ungeprüft) beantragt. Nach Absprache mit der Hochschulleitung werden diese Geräte bei Bewilligung durch die DFG aus der Grundausstattung beschafft.

	2005	2006	2007	2008
Für Sächliche Verwaltungsausgaben stehen als **Grundausstattung** voraussichtlich zur Verfügung:	37 500	37 500	37 500	37 500
Für Sächliche Verwaltungsausgaben werden als **Ergänzungsausstattung** beantragt (entspricht den Gesamtsummen "Sächliche Verwaltungsausgaben" in Übersicht 3.8):	3 500	5 500	1 500	1 500

(Alle Angaben in EUR)

Begründung zur Ergänzungsausstattung der Sächlichen Verwaltungsausgaben

(515) *Kleingeräte:*

einmalige Ausgaben 2005:	(5 000 EUR)	4 Rechner + 2 Monitore (zu beschaffen aus der Grundausstattung, s. o.)

(522) *Verbrauchsmittel:*

jährliche Ausgaben:	1 500 EUR	Büromaterialien, EDV-Zubehör und Software

(547) *Sonstiges:*

einmalige Ausgaben 2005:	2 000 EUR	Softwarelizenz für XEROX-Morphologie Werkzeuge
einmalige Ausgaben 2006:	4 000 EUR	Softwarelizenz für den Xerox Incremental Parser (XIP)

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

3.8.3. Investitionen (Geräte über EUR 10.000,- brutto und Fahrzeuge)

Linux-Cluster (EUR 25 000): Wie in Abschnitt 3.5.2 "Methoden und Arbeitsprogramm" dargelegt, ist die Arbeit mit großen Datenmengen ein zentraler Aspekt in allen Problembereichen des Projekts. In zum Ende dieser Projektphase durchgeführten Pilotstudien über das unüberwachte Lernen probabilistischer kontextfreier Grammatiken über sehr großen Korpora sowie über *k-nearest neighbor* Lernverfahren zeigte sich, dass in beiden Fällen aus paralleler Verarbeitung deutliche Laufzeitvorteile zu erwarten sind. Dabei profitieren die Formalismen allerdings von unterschiedlichen Ausprägungen paralleler Verarbeitung, die sich in unterschiedlichen Anforderungen an das beantragte Computersystem ausdrücken.

K-nearest neighbor Lernverfahren sind sehr rechenintensiv, arbeiten aber pro Durchlauf auf einer relativ kleinen Datenmenge. Die ideale Architektur zur parallelen Lö-

sung dieser Aufgabe ist ein *Cluster*, d.h. ein Zusammenschluss mehrerer Maschinen zu einem einzigen virtuellen Multiprozessorsystem. Prozess- und Speichermanagement wird dabei automatisch vom Cluster übernommen.

Beim Parsing großer Korpora ist eine andere Architektur von Vorteil. Hier ist wichtig, dass sehr große Datenmengen auf einzelne, dedizierte Maschinen verteilt werden und diese dann unabhängig voneinander geparst werden. Dies erfordert volle Kontrolle über die jeweilige Maschine, ihre Prozesslast und ihre Massenspeicher. Diese Architektur kann als *Rechnerverbund* bezeichnet werden.

Das beantragte Linux-Cluster soll in der Lage sein, beide Architekturen zu unterstützen. Dies ist möglich durch den Einsatz der quelloffenen Software *OpenMosix*, die mit einer speziellen Erweiterung des Linux-Systemkerns für das ausgeführte Programm vollständig transparent effizientes Prozess-Scheduling, Ressourcen- und Lastverteilung bereit stellt, und somit das System in den Cluster-Betriebsmodus versetzt.

Wird OpenMosix abgeschaltet, was ohne Rekonfigurationsaufwand bei laufendem System möglich ist, wechselt dieses in den Betriebsmodus als Rechnerverbund. In diesem Modus ist der volle programmtechnische Zugriff auf die Einzelmaschinen möglich.

Ein System, das den beschriebenen Anforderungen genügt, ist bisher an der Universität Tübingen nicht vorhanden.