Ergebnisbericht

Teilprojekt: A1

Thema: Repräsentation und Erschließung linguistischer Daten

Leiter: Prof. Dr. Erhard W. Hinrichs

Mitarbeiter:	Beata Kouchnir, M.A.	(wiss. Mitarb., seit 2/04)
	Martina Liepert	(wiss. Mitarb., 4/02-1/04)
	Frank H. Müller	(wiss. Mitarb., seit 4/02, davor wiss. HK)
	Dr. Heike Telljohann,	(wiss. Mitarb., 1/02-3/02, dann assoz. Mitarb. KIT[1])
	Julia Trushkina, M.A.	(wiss. HK, seit 3/04, davor assoz. Stipendiatin, Doktoranden-Stipendium der VW-Stiftung)
	Tylman Ule, M.A.	(wiss. Mitarb., seit 1/02)
	Holger Wunsch, M.A.	(assoz. Mitarb. KIT[1], 9/03-12/04)

Inhalt

[1]Kompetenzzentrum für Text- und Informationstechnologie

1. Kenntnisstand bei der Antragstellung und Ausgangsfragestellung

Wie in der ersten Projektphase bestand das Projektziel in der Förderphase 2002-2004 darin, Textkorpora des Deutschen so zu annotieren, dass die Korpusdaten für sprachwissenschaftliche Einzeluntersuchungen und computerlinguistische Anwendungen fruchtbar gemacht werden können. Der Schwerpunkt bei der Annotation lag auf komplexen syntaktischen Strukturen. Eine wesentliche Grundlage dafür bildet die robuste Annotation linguistischer Phänomene, die für die Makrostruktur deutscher Sätze von zentraler Bedeutung sind. Auf Basis einer bereits existierenden Chunkanalyse sollte 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, sollten die genannten linguistischen Annotationsprobleme so weit als möglich in einem *Shared-Task*-Szenario bearbeitet werden:

Im Sinne der Vergleichbarkeit der Resultate sollten die Methoden mit gleichen Ein- und Ausgaberepräsentationen arbeiten und auf gemeinsame linguistische Datenquellen zugreifen. Für die drei o.g. Untersuchungsbereiche wurden als übergeordnete und vergleichende Aufgaben (Shared Tasks) die Erkennung von topologischen Feldern und von grammatischen Funktionen spezifiziert:

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

Hinsichtlich der Gesamtthematik des SFB 441 zum Verhältnis von Empirie und Theorie in der Grammatikforschung wollte das Projekt einen Beitrag zur folgenden übergeordneten Fragestellung leisten: Welche Arten von linguistischen Informationen sind nötig, um zentrale grammatische Phänomene des Deutschen theorieneutral zu erschließen? Welche computerlinguistischen Analyseverfahren eignen sich für die korpuslinguistische Erschließung grammatischer Phänomene?

2. Angewandte Methoden

Die im Projekt angewandten Methoden beziehen sich auf die im Finanzierungsantrag genannten Arbeitsbereiche “Annotationsverfahren” und “quantitative und qualitative Evaluation”.

2.1. Annotationsverfahren

Für die in Abschnitt 1 beschriebenen Shared Tasks wurden mit regel- und datenbasierten Methoden jene Annotationsverfahren eingesetzt, die den Stand der Forschung insgesamt widerspiegeln.

Regelbasierte Verfahren: Grundlage der regelbasierten Verfahren bildeten die in der ersten Projektphase erarbeiteten Chunk-Analysen des Deutschen. Diese wurden für die Annotation komplexer syntaktischer Strukturen in zwei Richtungen erweitert:

1. *Finite-state (FS) Transduktoren*: Unter Verwendung der an der University of Edinburgh entwickelten TTT-Werkzeuge wurden Kaskaden von FS-Transduktoren entwickelt, die es erlauben, topologische Felder ebenso wie grammatische Funktionen zu annotieren. Als zusätzliche Wissensquellen wurden das Morphologie-System DMOR (Schiller 1995) und das elektronische Valenzwörterbuch IMSLex (Eckle-Kohler 1999) verwendet, die beide an der Universität Stuttgart erstellt wurden.
2. *Inkrementelle Dependenzgrammatische Annotation:* In Kooperation mit dem Xerox Research Center Europe (XRCE) wurde das *Xerox Incremental Parsing System* (XIP) für das Deutsche adaptiert und erweitert.

Datenbasierte Verfahren: Neben dem bereits in der ersten Projektphase eingesetzten symbolischen Lernverfahren des *memory-based learning* (MBL) wurden Experimente mit einem weiteren Lernverfahren, *Support Vector Machines* (SVMs), durchgeführt, wobei die Merkmals- und Parameterauswahl mit Hilfe von *genetischen Algorithmen* optimiert wurde. Außerdem wurden mit *probabilistischen kontextfreien Grammatiken* ein statistisches Verfahren eingesetzt, das zu den Standardverfahren datenbasierter Ansätze in der Computerlinguistik zu rechnen ist.

2.2. Quantitative und Qualitative Evaluation

Um die Güte der Annotation quantitativ überprüfen zu können, wurden Evaluationsverfahren für partiell annotierte Strukturen erarbeitet. Neben konstituentenbasierten Verfahren wurden dependenzorientierte Evaluationsmethoden für das Deutsche adaptiert und weiterentwickelt, die eine inkrementelle und theorieneutrale Evaluation komplexer grammatischer Phänomene ermöglichen.

SIMPX					Sätze
VF	LK	MF		RK	Felder
OD		ON			Funktionen
NX		NX			Phrasen
NX	VXFIN	NX	NX	VXINF	Chunks
Der Erklärung	war	eine etwa 45minütige Debatte	der Vollversammlung	vorausgegangen	Wörter
ART NN	VAFIN	ART ADV ADJA NN	ART NN	VVPP	PoS-Tags
nsm nsf gsf gsf dsf dsf gp0 asf	1s 3s	nsf asf nsf gsf dsf asf	nsm nsf gsf gsf dsf dsf gp0 asf		morpholog. Ambiguitäts-klasse

Abbildung 1: Annotationsschichten

Eine wesentliche Grundlage für die quantitative Evaluation großer Datenmengen bildet die Verfügbarkeit von qualitativ hochwertigen Testdaten im Sinne eines Goldstandards. Mit der *Tübinger Baumbank des Deutschen/Schriftsprache* (TüBa-D/Z) wurde eine syntaktisch annotierte Baumbank des Deutschen bereitgestellt, die durch die Annotation morphologischer Kategorien ergänzt und in das vom Projekt C1 entwickelte Annotationsformat TUSNELDA integriert wurde.

Neben der quantitativen Evaluation ist eine qualitative Evaluation erforderlich, die eine zielgerichtete Recherche von linguistisch relevanten Einzelphänomenen ermöglicht. Dazu wurde das in der ersten Projektphase entwickelte Recherche-Werkzeug VIQTORYA in seiner Funktionalität und Ausdrucksstärke erweitert. Mit der Integration der Baumbank-Daten in das TUSNELDA Format ergibt sich eine weitere Recherche-Möglichkeit über die vom Projekt C1 bereitgestellte XML-Datenbank und deren graphische Benutzeroberfläche.

3. Ergebnisse und ihre Bedeutung

Im Folgenden werden die Projektergebnisse anhand der im Finanzierungsantrag genannten Arbeitsbereiche “Annotationsverfahren” und “quantitative und qualitative Evaluation” dargestellt, wobei bei den Annotationsverfahren zwischen den in Abschnitt 2 genannten regel- und datenbasierten Verfahren unterschieden wird.

3.1. Shared Tasks

Abb. 1 stellt mehrere Schichten linguistischer Annotation anhand eines Beispiels dar. Wörter, Wortarten auf der Grundlage von Schiller, Teufel und Thielen (1995) (*PoS*)

und morphologische Ambiguitätsklassen bilden dabei die Eingabe, auf der alle im Projekt verwendeten Ansätze aufbauen. Folgende Shared Tasks wurden definiert, um die Leistungsfähigkeit unterschiedlicher Verfahren zur Annotation der dargestellten Informationen zu vergleichen, die zusammen die syntaktische Analyse deutscher Sätze ergeben:

Topologische Felder Das Modell der Topologischen Felder erlaubt es, die relativ freie Konstituentenfolge des Deutschen mit einem weithin akzeptierten und empirisch wohlfundierten Ansatz zu beschreiben (Höhle 1986). Komplexe Sätze werden in diesem Modell als Gefüge verbaler und non-verbaler Felder analysiert, wobei die finiten und infiniten Teile des Verbs zusammen die Satzklammer eines Satzes bilden (LK und VC in Abb. 1), und Komplemente und Adjunkte den Sätzen in Vor-, Mittel- und Nachfeld zugeordnet werden. Ein vollständig mit topologischen Feldern annotierter Satz schränkt also den Suchraum für grammatische Funktionen stark ein (*containment of ambiguity*). Dieser Vorteil wird ergänzt durch eine Regelhaftigkeit des Modells, die es erlaubt, topologische Felder mit robusten und effizienten Methoden des partiellen Parsens zu annotieren (*easy-first parsing*).

Grammatische Funktionen Während Konstituenten sich unabhängig von ihrer Distribution aufgrund ihrer inhärenten Eigenschaften z.B. als NP oder Teilsatz definieren, bezeichnen *grammatische Funktionen* (GFs) Relationen **zwischen** Konstituenten. Zudem drücken sie zumeist auch semantische Relationen aus und können somit Grundlage für eine semantische Analyse sein. Der Begriff *grammatische Funktion* orientiert sich 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. Die im Projekt annotierten GFs beschränken sich auf die Komplemente des Verbs und des Adjektivs, die in Tab. 1 aufgeführt werden.

Morphologische Disambiguierung Zusätzlich zu den beiden ursprünglich vorgesehenen Aufgabenfeldern ist im Verlauf des Projekts ein dritter Shared Task hinzugekommen, denn aufgrund des engen Zusammenhangs zwischen Kasusmarkierung und grammatischen Funktionen bildet die Disambiguierung von Kasusmerkmalen für die kasusrelevanten PoS eine wesentliche Voraussetzung für die korrekte Zuweisung grammatischer Funktionen. Da kasusmarkierte Wortformen im Deutschen häufig mehrfach ambig sind, ist eine morphologische Analyse notwendig, aber nicht hinreichend. Andererseits reflektieren die für das PoS-Tagging verwendeten Tagsets häufig keine feinkörnigen morphologischen Unterscheidungen einzelner PoS. So schließt das Stuttgart-Tübingen Tagset (STTS), das sich zum Standard für die PoS-Annotation des Deutschen entwickelt hat, z.B. keine Kasusinformationen für Nomina ein. Hierfür wurde ein erweitertes Tagset von Kategorien definiert, das die notwendigen morphologischen Unterscheidungen beinhaltet.

3.2. Regelbasierte Verfahren

3.2.1. Finite-State Transduktoren

Das FS-Verfahren hat an den folgenden Shared Tasks teilgenommen:

Topologische Felder FS-Transduktoren stellen einen robusten und effizienten Formalismus zur Annotation von Chunks dar. Eine Ausweitung dieses Formalismus auf die Annotation weiterer, höherer Ebenen der linguistischen Annotation war daher wünschenswert. Da topologische Felder und Teilsätze in ihrem Aufbau genau wie Chunks syntaktischen Restriktionen unterliegen, die sich aufgrund der relevanten PoS-Tags erfassen lassen, wurden diese Strukturen ebenfalls mit FS-Transduktoren annotiert. Da FS-Transduktoren keine rekursiven Strukturen erfassen können, topologische Felder und Sätze aber potenziell rekursiv sind, wurden rekursive Strukturen durch eine begrenzte Iteration der Transduktoren nachgebildet (Müller und Ule 2002). Dabei wurden zunächst topologische Felder, dann (Teil-)Sätze und dann Chunks annotiert. Somit handelt es sich bei dem angewandten Verfahren um ein gemischtes *Bottom-Up-Top-Down*-Verfahren. Dieses Verfahren hat sich unter dem Namen KaRoPars (Ule und Müller 2004) als Grundlage für die robuste Annotation eines großen Korpus des Deutschen bewährt, das der wissenschaftlichen Öffentlichkeit als *Tübinger Partiell Geparstes Korpus des Deutschen/Schriftsprache* (TüPP-D/Z) (Müller 2004*a*) zur Verfügung gestellt wurde. Für die topologischen Felder wurde die Konkurrenzfähigkeit von regelbasierten Verfahren mit auf Lernverfahren basierenden Ansätzen nachgewiesen (Veenstra, Müller und Ule 2002). Für die Erkennung von *Named Entities* (NEs), die ein besonderes Problem des Chunking darstellen, wurden die komplexen NPen in der TüBa-D/Z nach Token durchsucht, die diese komplexen Phrasen auslösen. So sind z.B. Titel Auslöser von Appositionen. Aus der Liste der Token wurde dann eine Liste von Lexemen erstellt, die zur Annotation der NEs innerhalb der kaskadierten FS-Transduktoren verwendet wurde.

Morphologische Disambiguierung Mit Hilfe der flachen Annotationsstruktur aus KaRoPars wird in Müller (2004*a*) die morphologische Ambiguitätsklasse der potenziellen grammatischen Funktionen robust durch ein einfaches Ranking-Verfahren reduziert. Dieses Verfahren basiert auf der Kongruenz aller Token in Noun Chunks (NCs) bezüglich Kasus, Numerus und Genus. Weiterhin wurde die flache Annotationsstruktur genutzt, um die Verb-Nominativ-Objekt-Kongruenz im Numerus abzugleichen und so die Anzahl möglicher Nominativobjekte einzuschränken. Das Problem, dass die morphologische Information der Token aus mehreren Merkmalen besteht und somit für einen FS-Ansatz schwerer zugänglich ist, wurde durch die Schaffung von Attribut-Wert-Kombinationen gelöst, die alle Merkmale vereinen. Weiterhin wurde der Ranking-Ansatz in den FS-Formalismus integriert.

Grammatische Funktionen FS-Ansätze werden i.A. für die Annotation flacher Strukturen verwendet. Die Anwendung dieser Methode auf die ‘tieferen’ Analyse-

strukturen der GFs stellt daher eine innovative Generalisierung dieser Methode dar. Während flache Strukturen jedoch einzig auf PoS-Tags aufbauen, benötigen GFs zu ihrer Annotation komplexere Datenstrukturen in Form von Information über Morphologie und Subkategorisierung (SubKat). Die Generalisierung war möglich, da die morphologischen Informationen in die flache Annotationstruktur integriert und SubKat-Rahmen in FS-Transduktoren umgewandelt werden konnten. Da die Distribution der grammatischen Funktionen jedoch im Gegensatz zur flachen Annotationsstruktur recht frei ist, bedarf es einer Entscheidungsstrategie für ambige Strukturen, die in Müller (2004*c*) detailliert dargelegt wird. Es werden für jeden Satz die zu dem SubKat-Rahmen des entsprechenden Verbs gehörenden Regeln in der umgekehrten Reihenfolge ihrer linguistischen Markiertheit aufgerufen (Vgl. Uszkoreit (1987)). Bei der Abfolge der Regel-Sätze für die entsprechenden SubKat-Rahmen gilt, dass sie in der Reihenfolge der Anzahl ihrer Komplemente (also *longest match*) angewendet werden. Tabelle 1 gibt die Ergebnisse für die Annotation grammatischer Funktionen auf der Grundlage der handannotierten PoS wider.

		Precision	Recall	$F_{\beta=1}$
gesamt		85.54%	79.65%	82.49
ON:	Objekt, Nominativ	91.36%	90.20%	90.77
OD:	Objekt, Dativ	75.95%	56.07%	64.52
OA:	Objekt, Akkusativ	81.99%	81.73%	81.86
OPP:	Objekt, präpositional	70.89%	44.94%	55.01
OS:	Objekt, Satz	73.21%	71.93%	72.57
PRED:	Prädikativ	83.50%	76.07%	79.61

Tabelle 1: Evaluation der Annotation grammatischer Funktionen

3.2.2. XIP

Als zweites regelbasiertes Verfahren wurde das vom Xerox Research Center Europe entwickelte *Xerox Incremental Parsing System* (XIP) für das Deutsche adaptiert und erweitert. Das XIP System unterscheidet drei Annotationsebenen: Wortklassenerkennung, Chunking und Dependenzanalyse.

Wortklassenanalyse und morphologische Desambiguierung Neben den notwendigen Änderungen auf der Ebene linguistischer Repräsentationen mussten auch Veränderungen am XIP System selbst vorgenommen werden. In Zusammenarbeit mit den Kollegen Ait-Moktar, Chanod und Roux wurde die Wortarten-Analysekomponente des XIP Systems um sog. *double-reduction rules* erweitert, die es erlauben, Kongruenzregeln zu spezifizieren, die für die phraseninterne, morphologische Desambiguierung von entscheidender Bedeutung sind (vgl. (Hinrichs und Trushkina 2003, Hinrichs und Trushkina 2004*a*)).

		Precision	Recall	FB1
gesamt		94.84%	93.27%	94.05
ON:	Objekt, Nominativ	95.94%	95.74%	95.84
OD:	Objekt, Dativ	69.35%	70.96%	70.15
OA:	Objekt, Akkusativ	92.66%	92.22%	92.44
OV:	Objekt, verbal	96.18%	95.65%	95.96
VPT:	abtrennbare Verbpartikel	96.87%	96.87%	96.87
PRED:	Prädikativ	75.28%	72.34%	73.78

Tabelle 2: Evaluation der Annotation grammatischer Funktionen in XIP

Die im erweiterten XIP System implementierte, regelbasierte Desambiguierungskomponente GRIP wurde anhand von Testdaten aus dem TüBa-D-Z Korpus evaluiert. Für 77.08 % aller NPen aus einem Testkorpus mit 57312 tokens und 1571 NPen konnte die GRIP Komponente mporphologisch eindeutige Lesarten zuweisen; nur 7.04 % der NPen behalten drei oder mehr Analysen.

Chunking Die Chunkingkomponente des XIP Systems dient v.a. als Vorverarbeitungsstufe für die Dependenzanalyse. Die Chunkrepräsentationen haben somit keinen eigenen Status. Dennoch hängt die Güte der nachfolgenden Dependenzanalyse natürlich indirekt von der Güte der vorgeschalteten Analyseebenen von morphologischer Desambiguierung und Chunkanalyse ab.

Grammatische Funktionen In XIP werden grammatische Funktionen im Dependenzparsing-Modul annotiert. Aufgrund der inkrementellen Struktur des Systems hängt das Dependenzmodul von der Annotation in den früheren Stadien ab. Dependenzregeln beinhalten Restriktionen über die Kategorien und morphologische Merkmale von Wörtern, deren syntaktischen Kontext und lineare Anordnung. Außerdem ist der Verwies auf Lemmata und elementare semantische Merkmale von Wörtern (z.B. *Zeit* und *Ort* für Nomina und *Performativa* für Verben) möglich, wodurch die Formulierung von differenzierteren Regeln ermöglicht wird. Tabelle 2 zeigt die Ergebnisse für die Annotation grammatischer Funktionen in XIP.

3.3. Datenbasierte Verfahren

3.3.1. Holistisches Memory-Based Parsing

Der Ansatz des holistischen *memory-based (MB) parsing* basiert auf den folgenden zwei Hypothesen:

1. Die Zuweisungen einzelner grammatischer Funktionen sind keine *unabhängigen* Entscheidungen und müssen daher in einem einzigen Schritt komplett vorgenommen werden.

2. Grammatische Funktionen können nur dann mit einem hohen Grad an Zuverlässigkeit annotiert werden, wenn der ganze Satz bei der Annotation berücksichtigt wird und nicht nur Satzausschnitte.

Dies bedeutet, dass komplette Baumstrukturen annotiert werden, der Parser ist jedoch darauf optimiert, grammatische Funktionen korrekt zu erkennen.

Aus der ersten Hypothese folgt, dass bei einem Klassifizierungsverfahren eine Klasse aus einem kompletten Syntaxbaum besteht und dass daher in den meisten Fällen der ausgewählte Syntaxbaum an den Eingabesatz angepasst werden muss. Aus der zweiten Hypothese folgt, dass eine flexible Anzahl von Merkmalen bei der Berechnung der Ähnlichkeit zur Anwendung kommt. Dann sind jedoch die gebräuchlichen Ähnlichkeitsmetriken nicht einsetzbar. Eine ausführliche Beschreibung dieser Metriken, sowie eine genauere Erklärung, warum sie hier nicht anwendbar sind, finden sich in der Dissertation von Sandra Kübler (Kübler 2002, Kapitel 2.2 und 5.1).

Unter Beibehaltung der beiden Hypothesen ist der Einsatz von MBL nicht möglich. Aus diesem Grund wurde ein neues Lernverfahren, auf MBL basierend, entwickelt. Dieses Verfahren wird hier kurz vorgestellt, eine ausführliche Beschreibung findet sich in der Dissertation von Sandra Kübler (Kübler 2002), kürzere Abhandlungen in Hinrichs, Kübler, Müller und Ule (2002), Kübler (2003) und Kübler (2004).

In der aktuellen Phase wurde im wortbasierten Modul die Ähnlichkeitsdefinition so modifiziert, dass eine größere Flexibilität erreicht wurde: Anstatt eine komplette Übereinstimmung in der Wortfolge zu verlangen, wurde das "Überspringen" von einzelnen Wörtern oder Chunks erlaubt. Jedes Überspringen ist jedoch mit Erhöhung eines Gewichts verbunden, die Analyse mit dem niedrigsten Gewicht wird bevorzugt weiterverfolgt. Auch das *backing-off*-Modul wurde flexibler und leistungsfähiger gestaltet; dieses wird statt der bei MBL üblichen Gewichtung von Merkmalen verwendet. Bereits implementiert waren Strategien, die auf der PoS-Ebene suchen. In der aktuellen Phase wurde außerdem eine Strategie implementiert, die die Suche auf die Ebene der Chunk-Sequenzen verlagert. D.h. alle Trainingssätze werden beim Training chunk-geparst, und die Chunk-Sequenzen innerhalb von Simplex-Sätzen werden in der Instanzenbasis abgelegt. Der Eingabesatz wird ebenfalls chunk-geparst, und aufgrund der so ermittelten Chunk-Sequenz wird der ähnlichste Satz in der Instanzenbasis ermittelt. In dem so gefundenen Syntaxbaum müssen dann noch die Strukturen innerhalb der Phasen an die Eingabe angepasst werden.

Dieses Modul hat sich als sehr erfolgreich herausgestellt, wie Tabelle 3 deutlich macht. Die Zuweisung von grammatischen Funktionen ist geringfügig schlechter als beim Gesamtsystem, die Erkennung von Konstituenten verbessert sich dagegen.

	Gesamtsystem	nur *backing-off*
Labeled Recall (Konstituenten)	82.45%	82.95%
Labeled Precision (Konstituenten)	87.25%	87.96%
$F_{\beta=1}$	84.78	85.38
Labeled Recall (+ gramm. Funktionen)	71.72%	70.52%
Labeled Precision (+ gramm. Funktionen)	75.79%	74.73%
$F_{\beta=1}$	73.70	72.56
Recall von angebundenen gramm. Funktionen	95.31%	94.63%
Precision von angebundenen gramm. Funktionen	95.21%	94.51%
$F_{\beta=1}$	95.26	94.57

Tabelle 3: Ergebnisse des holistischen Ansatzes

3.3.2. Inkrementelle Verfahren

Memory-Based Learning (MBL) Maschinelle Lernverfahren wurden parallel zum oben genannten holistischen Ansatz auch als Klassifikationsverfahren zur inkrementellen Annotation verwendet und für die folgenden Shared Tasks eingesetzt:

a) Topologische Felder In Anlehnung an die Annotation von PoS wurde ein *Fenster-Ansatz* verwendet, bei dem einem Fokus-Wort als potenziellem Teil der Satzklammer eine Klasse zugewiesen wird. Als Eingabemerkmal werden die PoS der zwei vorausgehenden Wörter, des Fokus-Worts und des folgenden Wortes verwendet (Veenstra et al. 2002). Zielklassen sind IOB-Tags, also Klassen, die bestimmen, ob das Fokus-Wort am Anfang, außerhalb oder zu Beginn aufeinanderfolgender Satzklammern steht. Die erfolgreichste Gewichtung durch *Information Gain* zeigt für Wortformen als optionale Eingabemerkmale geringe Werte an, was auf zu geringen Umfang dieses Typs von Trainingsdaten hinweist.[2] Trotz des beschränkten Kontexts erwies sich der gewählte Ansatz mit $F_{\beta=1} = 0.97$ als sehr erfolgreich (Veenstra et al. 2002).

b) Grammatische Funktionen Da unzverlässige externe Informationen (z.B. Morphologie, Semantik) oft die Leistung eines Lernalgorithmus beeinträchtigen, werden für die Annotation grammatischer Funktionen mit MBL ausschließlich lexikalische und syntaktische Informationen aus der Baumbank verwendet (Kouchnir 2004). Das Modell enthält Merkmale sowohl über die tiefe (Konstituente, lexikalischer Kopf) als auch die flache (topologische Felder, Satztyp) Phrasenstruktur; ausserdem verwendet es Informationen über den syntaktischen (finite Verb) und semantischen (Vollverb) Kopf des Satzes, und den linearen Kontext. Tabelle 4 zeigt die Ergebnisse für die Annotation grammatischer Funktionen für Komplemente und Adjunkte.

[2] Zum Zeitpunkt der Untersuchung enthielt die TüBa-D/Z weniger als 100 000 Wörter.

Funktion	Precision (%)	Recall (%)	$F_{\beta=1}$
Gesamt	76.95	77.71	76.69
ON	84.38	90.28	87.23
OA	77.19	80.08	78.60
OD	88.23	33.64	48.71
OS	71.58	79.70	75.42
OPP	64.11	60.18	62.09
PRED	78.64	72.27	75.32
FOPP	62.75	32.48	42.80
MOD	81.22	81.03	81.12
ON-MOD	43.32	20.59	27.91
OA-MOD	46.41	35.52	40.24
V-MOD	71.55	81.61	76.25

Tabelle 4: Ergebnisse für die Annotation grammatischer Funktionen mit MBL

Support Vector Machines Neben MBL wurden SVMs als Klassifikationsverfahren für die Annotation topologischer Felder verwendet, da SVMs über polynomiale Kernel-Funktionen optimale Kombinationen von Merkmalen bilden können, und sie sich im Vergleich mit anderen Lernverfahren oft als mindestens ebenbürtig erwiesen haben (Kudo und Matsumoto 2001, für Chunking). Jedoch ist bei ihrer Anwendung auf Probleme der Sprachverarbeitung noch nicht geklärt, wie die oft hoch asymmetrische Verteilung von Klassen adäquat in die Trainingsparameter einfließen soll, insbesondere bei getrennter Gewichtung positiver und negativer Instanzen. Ebenso ist die optimale Behandlung von nicht-binären Zielklassen noch nicht abschließend geklärt. Letzteres Problem wurde durch die Verwendung mehrerer Instanzen von Klassifikatoren gelöst, die sich zwischen jeweils zwei Zielklassen entscheiden und die über einfache Mehrheitswahl kombiniert werden. Das Problem der Gewichtung wurde durch die Anwendung genetischer Algorithmen gelöst, die mittels Mutation und Rekombination erfolgreicher Parametersätze die Zielfunktion der Annotationsgüte optimieren (Liepert 2003).

3.3.3. Probabilistische Kontextfreie Grammatiken

Probabilistische kontextfreie Grammatiken (PCFGs) finden als Bindeglied zwischen daten- und regelbasierten Ansätzen Anwendung. PCFGs haben an den folgenden Shared Tasks teilgenommen:

Topologische Felder Die strenge Regelmäßigkeit des Aufbaus komplexer Satzstrukturen aus topologischen Feldern sowie ihre fast ausschließliche Abhängigkeit von PoS legt nahe, dass sich unmodifizierte PCFGs zur Annotation aller Teile der topologischen-Felder-Struktur von Sätzen eignen. Als Hauptproblem stellte sich da-

bei die sehr starke Annahme der Kontextfreiheit in Zusammenhang mit der Wahl von Struktur und Kategorien in TüBa-D/Z heraus. Zur Verbesserung der Annotationsgüte wurden daher Unterklassen von Kategorien gebildet, wo Kategorien von Knoten in bestimmten Kontexten komplementäre Verteilungen von Tochterknoten aufweisen. Beispielsweise enthalten Verbalkomplexe immer und ausschließlich in Verb-Letzt-Sätzen finite Verben (Veenstra et al. 2002). Diese zunächst von Hand eingeführten Optimierungen an PCFGs zur Verbesserung der Annotationsgüte bei der Annotation der Satzklammer haben zu der Frage geführt, welche Optimierungen der Struktur einer Grammatik zur optimalen Repräsentation der gewünschten Zielstruktur im Sinne einer bestimmten Annotationsmethode führen. Neben vollständig komplementären Verteilungen der Produktionen von PCFGs sind ebenso Fälle zu erwarten, in denen ein Knoten zwar über alle Vorkommen ähnliche Mengen von Tochterknoten aufweist, für die sich jedoch abhängig vom weiteren Kontext signifikante Unterschiede in der Häufigkeitsverteilung ergeben. So erscheinen Adjektivphrasen (APen) als direkte Töchter von Feldern ebenso wie als Töchter von NPen. Die Verteilung ihrer Produktionen ist dabei stark verschieden: sind im Mittelfeld 0,007% aller Töchter von APen attributive Adjektive, so sind es 73% unter NPen innerhalb von PPen. Zur automatischen Optimierung von Repräsentationen solcher Verteilungspräferenzen wurde erfolgreich *Directed Treebank Refinement* entwickelt (Ule 2003).

Disambiguierung von Morphologie PCFGs wurden mit Erfolg als datenbasierte Alternative zum in Abschnitt 3.2.2 vorgestellten regelbasierten Ansatz für die morphologische Disambiguierung eingesetzt. Hinrichs und Trushkina (2003) zeigen, dass PCFG-Modelle für das Deutsche bessere Ergebnisse mit dem in Abschnitt 3.2.2 beschriebenen erweiterten STTS-Tagset liefern als n-gramm-basierte Ansätze, wie sie von Tufiş (2000) bzw. von Dienes und Oravecz (2000) für das PoS-Tagging mit großen Tagsets vorgeschlagen worden sind. Wie für das syntaktische Parsing mit PCFGs ist auch für die Güte der morphologischen Disambiguierung entscheidend, dass das verwendete statistische Modell die den PCFGs zugrunde liegenden Unabhängigkeitsannahmen durch entsprechende Baumtransformationen der verwendeten Trainingsdaten in geeigneter Weise abfedert.

Die mit einem reinen PCFG-Modell erzielbaren Ergebnisse lassen durch die Vorschaltung eines regelbasierten Moduls weiter verbessern. Trushkina und Hinrichs (2004) beschreiben ein derartiges hybrides Modell und erzielen signifikante Verbesserungen im Vergleich zur Performanz reiner statistischer und regelbasierter Modelle.

3.4. Evaluation, Annotation und Recherche

3.4.1. Evaluation durch Konstituentenstruktur und Dependenzen

Die Annotation verschiedener Phänomenbereiche mit Hilfe unterschiedlicher Methoden hat ergeben, dass globale Evaluationsmetriken wie $F_{\beta=1}$ über alle Klassen anno-

tierter Konstituenten selbst dann ähnliche Werte zeigen, wenn einzelne Konstituententypen von den Methoden verschieden gut verarbeitet werden. So weichen PCFG und FSA bei der Annotation topologischer Felder in Teilen der Zielannotation weit mehr voneinander ab als wenn über alle Zielklassen evaluiert wird (Veenstra et al. 2002). Daher wurde die TüBa-D/Z mit Informationen zu Dependenz-Beziehungen angereichert, so dass in einer gemeinsamen Datenstruktur sowohl Konstituenten als auch Dependenzrelationen vereinigt sind (Ule und Kübler 2004) und Vergleiche von Anbindungen die wesentlichen Unterschiede zeigen. Als Voraussetzung wurde zunächst gezeigt, dass alle hierfür notwendigen Informationen bereits in der Konstituentenstruktur von TüBa-D/Z kodiert sind (Kübler und Telljohann 2002). Abb. 2 zeigt einen auf Relationen basierenden Vergleich der vollautomatischen Annotation der TüPP-D/Z mit der handannotierten TüBa-D/Z. Die Label der Dependenzrelationen repräsentieren dabei die komplexe TopF-Struktur; in der automatischen Annotation fehlende Informationen sind gestrichelt gegeben, und solche, die nicht in der Gold-Annotation vorhanden sind, befinden sich unterhalb des Textes. Im Beispiel ist also die komplexe Satzstruktur korrekt annotiert und einzig der Präpositional-Chunk zu hoch angebunden.

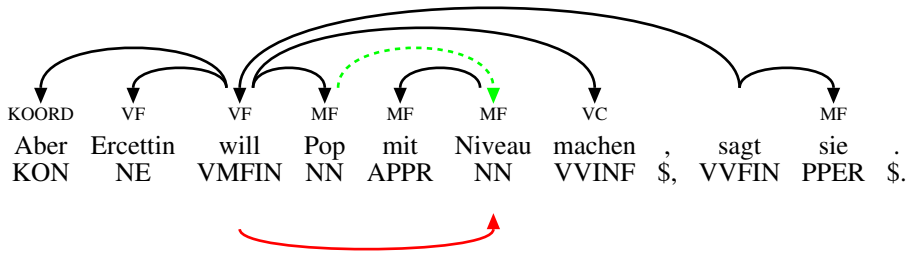

Abbildung 2: Vergleich von TüPP-D/Z und TüBa-D/Z mittels Dependenzbeziehungen

Fehler in der zum Vergleich herangezogenen Gold-Annotation stören insbesondere die Evaluation. Daher wurde eine Methode entwickelt, die mit Hilfe eines statistischen Tests unerwartete und seltene Teilstrukturen in Korpora findet. Diese Methode wurde erfolgreich zum Finden von Annotationsfehlern in Korpora verschiedener Sprachen, Größe und Bearbeitungszustände eingesetzt (Ule und Simov 2004).

3.4.2. Annotation

Die in Kooperation mit dem Kompetenzzentrum für Text- und Informationstechnologie (KIT) entwickelte TüBa-D/Z ist ein syntaktisch annotiertes Korpus, das auf Textmaterial aus der Wissenschafts-CD der Zeitung “die tageszeitung” (taz) basiert. Die bisher annotierten Zeitungstexte stammen aus den Ausgaben vom 3. bis 7. Mai 1999. Seit Dezember 2003 ist die TüBa-D/Z zusammen mit einem detaillierten Stylebook (Telljohann, Hinrichs und Kübler 2003) als Ressource für Forschungs- und Entwicklungszwecke unter `http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml` erhältlich. Die Baumbank umfasst derzeit 15 260 Sätze (266 411 Wörter).

Die syntaktische Struktur der Baumbank wird auf vier Ebenen beschrieben: der lexikalischen Ebene, der phrasalen Ebene, der Ebene der topologischen Felder und der Satzebene. Das Konzept der topologischen Felder ermöglicht die Beschreibung der Wortstellungsregularitäten des Deutschen (relativ hoher Grad an Nichtkonfigurationalität) und begünstigt eine Annotation, die weder kreuzende Kanten noch Spuren verwendet. Nicht-adjazente Relationen werden durch spezifische Kantenlabel dargestellt. Darüber hinaus werden Secondary Edge Label eingeführt, um folgende Phänomene zu beschreiben: Dependenzrelationen im Verbkomplex, Modifikation von phraseninternen NPen, Auflösung von Long-Distance-Beziehungen unter Modifikatoren sowie Dependenzen in Kontrollverbkonstruktionen. Der Vergleich der TüBa-D/Z mit der unabhängig entwickelten TIGER Baumbank macht grundsätzliche Unterschiede in der Behandlung der freien Wortstellung des Deutschen und der diskontinuierlichen Konstituenten deutlich (Telljohann, Hinrichs und Kübler 2004).

Die Bearbeitung der Shared Tasks im Projekt A1 erforderte die Erweiterung der Annotation in dreierlei Hinsicht: 1) NEs, 2) NCs und 3) Einbindung der Morphologie.

Um NEs in der Baumbank zu markieren, wurden ein zusätzliches Knotenlabel und ein Secondary Edge Label definiert. Das Knotenlabel kennzeichnet, dass der darunter liegende Knoten eine NE repräsentiert. Das Secondary Edge Label wird verwendet, wenn die Einfügung des Knotenlabels eine Veränderung der syntaktischen Struktur zur Folge hätte, d.h. dieses Label gibt Informationen über die Relation zwischen zwei Teilen einer NE innerhalb einer komplexen Phrase. Darüber hinaus wurden Informationen über NCs in nominale Konstituenten integriert, d.h. nominale Knoten werden durch einen spezifischen NC-Knoten ersetzt. Bezüglich der Morphologie wurden für die ersten 7 000 Sätze die morphologischen Ambiguitätsklassen halbautomatisch disambiguiert und in die Baumbank integriert.

3.4.3. Recherche

Um in einem Korpus nach linguistischen Phänomenen suchen zu können, sind zwei Voraussetzungen notwendig: ein Korpus, das diese Phänomene kodiert, sowie ein Query-Tool mit einer Query-Sprache, die diese Phänomene beschreiben kann.

Korpora, die für syntaktische Fragestellungen die erste Bedingung erfüllen, sind im SFB 441 vorhanden: zum einen die in A1 entwickelte TüBa-D/Z und die in Verbmobil entwickelte *Tübinger Baumbank des Deutschen/Spontansprache* (TüBa-D/S), zum anderen die in A3 entwickelte Sammlung suboptimaler Strukturen oder das in B11 annotierte Textkorpus des Tibetischen.

Zur Erfüllung der zweiten Bedingung wurde in der Förderphase 1999-2001 der Prototyp des Query-Tools VIQTORYA (A **Vi**sual **Q**uery **T**ool f**or** S**y**ntactically **A**nnotated Corpora) entwickelt und implementiert.

In der aktuellen Phase lag das Hauptaugenmerk bei der Weiterentwicklung von VIQTORYA auf den Anforderungen, die durch die Verwendung von VIQTORYA für die qualitative Evaluation der Annotationsverfahren entstanden, ebenso wie auf den Bedürfnissen der anderen Projekte im SFB 441. Es wurden vor allem Erweiterungen in drei Hauptbereichen angegangen: Erweiterung der Query-Sprache um Disjunktionen und um globale Negation (Schurtz 2002), Erweiterung der repräsentierten linguistischen Informationen, Verbesserung der Korpusverwaltung.

Die Erweiterung der Query-Sprache um Disjunktionen und um Negation erweiterte die Ausdruckfähigkeit der Query-Sprache bedeutend. Nun ist es z.B. auch möglich, alle Sätze zu suchen, die das ON nicht im Vorfeld (VF) haben, ohne alle Möglichkeiten aufzulisten. Bei der Erweiterung der repräsentierten linguistischen Informationen wurden zwei weitere Informationskategorien, die im *Annotate*-Format vorliegen, in VIQTORYA integriert: morphologische Annotation und Relationen, die über eine reine Baumstruktur hinausgehen (d.h. Secondary Edges). Beide Kategorien sind nun über die Query-Sprache abfragbar. Die Korpusverwaltung wurde so erweitert, dass zum einen verschiedene Korpora mit Teilkorpora geladen werden können und zum anderen verschiedene Anfragen mit Kommentaren gespeichert werden können. Alle Erweiterungen bedingten auch eine Anpassung der Benutzerschnittstelle und des Datenbankformats.

4. Vergleich mit Arbeiten außerhalb des Sonderforschungsbereichs und Reaktionen der wissenschaftlichen Öffentlichkeit auf die eigenen Arbeiten

4.1. Regelbasierte Verfahren

Unter Verwendung des FS-Formalismus werden GFs für das Türkische von Oflazer (2003), für das Französische von Aït-Mokhtar, Chanod und Roux (2002) und für das Deutsche von Schiehlen (2003) annotiert. Oflazer (2003) beschränkt sich ebenso wie das Projekt A1 auf Komplemente, evaluiert sein Verfahren jedoch lediglich auf 200 Sätzen, von denen 30 bereits im Trainings-Korpus enthalten sind. Aufgrund der nicht vorhandenen relevanten Nominalflexion im Französischen beschränken sich Aït-Mokhtar et al. (2002) auf die im Französischen auch stärker restringierten Stellungseigenschaften der GFs, die zusätzlich einen Verzicht auf die Verwendung von SubKat-Rahmen erlauben. Bei der Evaluation beschränken sich Aït-Mokhtar et al. (2002) auf die beiden Kategorien Subjekt und Objekt, während das Projekt A1 sechs GFs einbezieht. Schiehlen (2003) verwendet ebenfalls einen inkrementellen FS-Ansatz, der jedoch bei der Behandlung von Anbindungsambiguitäten auf Techniken constraintbasierter Grammatikformalismen zur Behandlung von Unterspe-

zifiziertheit zurückgreift und somit keinen reinen FS-Ansatz verfolgt. Bezüglich der auch von uns annotierten GFs erweist sich unser Parser als konkurrenzfähig. Ein weiterer Dependenzgrammatik-Ansatz bezogen auf das Deutsche findet sich bei Duchier (1999), Duchier (2000). Hier liegt der Schwerpunkt vorrangig auf der Entwicklung einer Parsing-Architektur im Rahmen des *Constraint Logic Programming*-Paradigma.

Klatt (2002) hat ein weiteres hybrides Taggingsystem für das Deutsche entwickelt. Er verwendet jedoch das STTS Tagset und beschränkt sich somit auf reines POS Tagging. In Trushkina und Hinrichs (2004) wird das STTS hingegen um morpho-syntaktische Merkmale erweitert. Das erweiterte Tagset umfasst 718 Tags, so dass die Taggingaufgabe weit über einfaches POS Tagging hinausgeht.

4.2. Datenbasierte Verfahren

Dubey und Keller (2003) stellen den ersten probabilistischen Parser für die volle syntaktische Annotation des Deutschen vor. Sie passen zur Verbesserung der Annotationsgüte einen Parser, der für das Englische entwickelt wurde, an die flachere Kodierung der Annotation des verwendeten Negra-Korpus an. Frank, Becker, Crysmann, Kiefer und Schäfer (2003) nutzen einen probabilistischen Topologische-Felder-Parser zur Vorstrukturierung der Eingabe eines mächtigeren HPSG-basierten Parsers. Sie erreichen damit einen beträchtlichen Geschwindigkeitszuwachs neben erhöhter Robustheit, evaluieren allerdings ebenfalls auf automatisch aus der Negra-Baumbank generierten Daten, wodurch ein direkter Vergleich nicht möglich ist. Zur Steigerung der Annotationsgüte nehmen sie wie Hinrichs und Trushkina (2003) oder Klein und Manning (2003) manuell Änderungen an der Repräsentation linguistischer Informationen vor. Letztere zeigen, dass die Annotationsgüte nichtlexikalisierter PCFGs durch solche Änderungen in den Bereich lexikalisierter PCFGs vorstößt. Eine entsprechende Optimierung der Repräsentation syntaktischer Annotation wurde im Projekt A1 erfolgreich automatisch durchgeführt (Ule 2003).

Wie Kübler (2002) verfolgt auch Streiter (2001) einen holistischen Ansatz des MB Parsing. Streiter konzentriert sich auf das Parsen chinesischer Daten. Die von ihm verwendeten Schlüsselwörter zum Retrieval der *k nearest neighbors* basieren jedoch nicht auf linguistischen Prinzipien und sind daher sprach- und implementierungsspezifisch. Der ähnlichste Baum wird in einem zweiten Schritt durch eine gewichtete Alignierung ermittelt. Buchholz (2002) verwendet einen inkrementellen MBL-Ansatz zur Ermittlung grammatischer Funktionen in der Penn Treebank, wobei dort im Unterschied zu TüBa-D/Z nur solche Komplemente und Adjunkte ausdrücklich markiert sind, die nicht über die Satzstellung erschlossen werden können.

4.3. Evaluation, Annotation und Recherche

Die Evaluation von Parsern ist derzeit ein wichtiger Bereich in der Computerlin-

guistik. Das gesteigerte Interesse zeigt sich z.B. an dem LREC-Workshop *Beyond PARSEVAL–Towards Improved Evaluation Measures for Parsing Systems* (Carroll 2002). Dort wurden neben dependenzorientierten auch hierarchische Ansätze diskutiert, die es erlauben, auf verschiedenen Ebenen von Generalität zu evaluieren.

Neben der TüBa-D/Z existiert mit TIGER (Brants, Dipper, Hansen, Lezius und Smith 2002) eine weitere Baumbank für das Deutsche. Diese Baumbank umfasst momentan 40 000 Sätze, das Annotationsschema ist gekennzeichnet durch die Verwendung von kreuzenden Kanten, eine erweiterte Menge von grammatischen Funktionen und eine sehr flache Annotation der phraseninternen Struktur. Die *Prague Dependency Treebank* (Hajic und Uresova 2003) ist eine große Baumbank für das Tschechische. Die Annotation ist dependenzorientiert, wobei zwischen verschiedenen Ebenen der Annotation unterschieden wird: der morphologischen Ebene, der analytischen Ebene (oberflächen-dependenzielle Syntax) und der tektogrammatischen Ebene (syntaktisch-semantische Struktur). Die Annotationen auf den verschiedenen Ebenen sind untereinander verbunden.

Lezius (2002) stellt das Suchwerkzeug TIGERSearch vor, das es erlaubt, Korpora für die Recherche aufzubereiten und auf sie graphisch definierte Abfragen anzuwenden. Das im Projekt A1 entwickelte Werkzeug VIQTORYA nutzt dagegen eine menügesteuerte Eingabe von Anfragen. Die Entwicklung von TIGERSearch schloss die Definition einer Abfragesprache und eines allgemeinen Repräsentationsformat annotierter Korpora ein, die sich zu einer Korpusbeschreibungssprache ergänzen. Im Gegensatz zu VIQTORYA, das eine darunter liegende Datenbank (mySQL) verwendet, nutzt TIGERSearch selbst generierte Indizes.

5. Offene Fragen

Ausweitung der grammatischen Funktionen auf Adjunkte In der derzeitigen Projektphase lag der Schwerpunkt bei der Annotation grammatischer Funktionen auf den Komplementen des Verbs. Es stellt sich nun die Frage, ob auch Adjunkte mit den gleichen bzw. ähnlichen Methoden zuverlässig annotiert werden können. Es muss angenommen werden, dass die Arten linguistischer Information, die eine erfolgreiche automatische Analyse von Komplementen erlauben sich nicht direkt auf die Annotation von Adjunkten übertragen lassen. Gleichzeitig bilden die bereits annotierten Komplemente eine neue Klasse von linguistischer Information, der für die Annotation von Adjunkten eine wichtige Grundlage bilden kann.

Behandlung von Parataxe, Hypotaxe und komplexen Koordinationen In der jetzigen Phase lag der Schwerpunkt auf der automatischen Annotation komplexer syntaktischer Strukturen, speziell von komplexen Phrasen, topologischen Feldern, und grammatischen Funktionen. Dabei haben sich die eingesetzten Annotationsverfahren

auf die Analyse von Simplexsätzen[3] beschränkt.

Es musste jedoch ungeklärt bleiben, ob dieselben Annotationsverfahren und die für die Erkennung von Simplexsätzen bisher verwendeten linguistischen Informationen ausreichen, um auch komplexe Satzgefüge, insbesondere die Phänomene Hypotaxe, Parataxe und komplexe Koordinationen, annotieren zu können.

Anaphernresolution Werden komplexe Satzgefüge in den Blick genommen, stellen anaphorische Beziehungen zwischen Satzgliedern und die Auflösung anaphorischer Referenz ein zentrales Annotationsproblem dar.

Eine Behandlung der Anaphernresolution für das Deutsche[4] ist jedoch nur dann erfolgversprechend, wenn eine zuverlässige und effiziente Analyse der grammatischen Funktionen geleistet werden kann, wie sie in der derzeitigen Phase in A1 entwickelt wurde.

Mit der Anaphernresolution ergibt sich ein Problembereich, der sich vor allem durch zwei Charakteristika von den bisher in A1 untersuchten Problembereichen unterscheidet: Zum einen wird wie beim vorherigen Punkt die Grenze der Simplex-Sätze überschritten. Zum anderen handelt es sich hier um einen Problembereich, der im Bereich der Syntax-Semantik-Schnittstelle angesiedelt ist.

Es ergibt sich jedoch auch hier wieder die Frage, welche Typen linguistischer Information nötig sind, um Anaphern auflösen zu können.

Hybride und nicht überwachte Lernmethoden Die Annotation der oben aufgeführten Phänomene führt sehr schnell zum Problem der *data sparseness*, d.h. zu kleiner Trainingskorpora. Als möglicher Lösungsweg bieten sich hybride Ansätze an, bei denen verschiedene Verfahren maschinellen Lernens verknüpft werden, um somit die Menge an verwertbarer Information zu vergrößern. In diesem Bereich soll der holistische MBL Ansatz zu einem flexibleren *k*-nearest neighbor Ansatz erweitert werden.

Ein anderer Ansatz besteht in nicht oder minimal überwachten Lernverfahren. Zum Ende dieser Projektphase wurde bereits ein erster Ansatz in dieser Richtung verfolgt, der als Basisformalismus probabilistische kontextfreie Grammatiken (*Probabilistic Context-Free Grammars, PCFGs*) verwendet. Eine überwacht trainierte PCFG soll mithilfe unüberwachter Lernverfahren (Baum-Welch Algorithmus zur iterativen Ermittlung der probabilistischen Parameter einer PCFG) verbessert werden. Hierbei kommt das in dieser Projektphase aufgebaute große TüPP-D/Z Korpus zum Einsatz, als Goldstandard wird TüBa-D/Z verwendet.

[3] Der Begriff *Simplexsatz* entspricht der englischen Bezeichnung von *simplex clause* und soll Sätze bezeichen, die keine satzwertigen Einbettungen haben.

[4] Bisher liegen computerlinguistische Ansätze zur Anaphernresolution größtenteils für die englische Sprache vor.

Die Beschränkungen von PCFGs zur Modellierung sprachlicher Phänomene, wie auch die Grenzen unüberwachten Lernens sind bekannt. Daher wäre es von besonderem Interesse, zu klären, inwieweit andere Ansätze, oder deren Kombination, weitergehende Verbesserungen erzielen können.

6. Veröffentlichungen und Manuskripte

6.1. Veröffentlichungen von Mitarbeitern

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. und K. Simov (Hrsg.) (2002): *Proceedings of the First Workshop on Treebanks and Linguistic Theories, TLT 2002*, Sozopol, Bulgaria.

Hinrichs, E.W. und J. Trushkina (2002*a*): „Forging Agreement: Morphological Disambiguation of Noun Phrases“, in *Proceedings of TLT 2002*, Sozopol, Bulgarien, S. 78–95.

Hinrichs, E.W. und J. Trushkina (2002*b*): „Getting a Grip on Morphological Disambiguation“, in *Tagungsband der KONVENS 2002*, Saarbrücken, S. 59–66.

Hinrichs, E.W. und J. Trushkina (2003): „N-gram and PCFG models for morpho-syntactic tagging of German“, in *Proceedings of TLT 2003*, Växjö, Schweden, S. 81–92.

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.

Hinrichs, E.W. und J. Trushkina (2004*c*): „Treebank Transformations for Performance Optimizations of a PCFG-based Tagger“, in *Pre-Proceedings of the International Conference on Linguistic Evidence*, Tübingen, S. 66–70.

Kübler, S. (2003): „Parsing Without Grammar – Using Complete Trees Instead“, in *Proceedings of RANLP 2003*, Borovets, Bulgarien.

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)

Kübler, S. und H. Telljohann (2002): „Towards a Dependency-Based Evaluation for Partial Parsing“, in *Proceedings of the LREC 2002-Workshop Beyond PARSEVAL*, Las Palmas, Spanien, S. 9–16.

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. 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.

Müller, F.H. und T. Ule (2003): „On the nature, annotation and use of shallow parsing structures“, *in* L. Cyrus, H. Feddes, F. Schumacher und P. Steiner, (Hrsg.), *Sprache zwischen Theorie und Technologie. Festschrift für Wolf Paprotté zum 60. Geburtstag*, Deutscher Universitäts-Verlag, Wiesbaden, S. 199–209.

Nivre, J. und E. Hinrichs (Hrsg.) (2003): *Proceedings of the Second Workshop on Treebanks and Linguistic Theories, TLT 2003*, Växjö, Schweden.

Steiner, I. und L. Kallmeyer (2002): „VIQTORYA – A Visual Query Tool for Syntactically Annotated Corpora“, in *Proceedings of LREC 2002*, Las Palmas, Spain.

Telljohann, H., E.W. Hinrichs und S. Kübler (2004): „The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone“, in *Proceedings of LREC 2004*, Lissabon, Portugal.

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. und E. W. Hinrichs (2004): „Linguistische Annotation“, *in* H. Lobin und L. Lemnitzer (Hrsg.), *Texttechnologie*, Stauffenburg, Tübingen, S. 217–244.

Ule, T. und F.H. Müller (2004): „KaRoPars: Ein System zur linguistischen Annotation großer Text-Korpora des Deutschen“, *in* A. Mehler und H. Lobin, (Hrsg.), *Automatische Textanalyse*, VS Verlag, Opladen, S. 185–202.

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

Ule, T. und K. Simov (2004): „Unexpected Productions May Well be Errors“, in *Proceedings of LREC 2004*, Lissabon, Portugal.

Ule, T. und S. Kübler (2004): „From Constituent Structure to Dependencies, and Back“, in *Pre-Proceedings of the International Conference on Linguistic Evidence*, Tübingen.

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

6.2. Qualifikationsarbeiten

Kübler, S. (2002): *Memory-Based Parsing of a German Corpus*, Dissertation, Universität Tübingen. Version vom 3. November 2002. Überarbeitete Version angenommen in der Buchreihe “Natural Language Processing”, Amsterdam: John Benjamins.

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

Schurtz, T. (2002): *Erweiterung des VIQTORYA-System um Disjunktionen*, Studienarbeit, Universität Tübingen.
URL: *http://www.sfb441.uni-tuebingen.de/a1/Publikationen/studienarbeit.pdf*

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

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

6.3. Manuskripte

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

Müller, F.H. (2004*b*): *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*, SfS, Universität Tübingen.
URL: *http://www.sfs.uni-tuebingen.de/tupp/dz/stylebook.pdf*

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*

Ule, T. (2004*b*): *Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)*, SfS, Universität Tübingen.
URL: *http://www.sfs.uni-tuebingen.de/tupp/dz/markupmanual.pdf*

7. Aktivitäten: Tagungen, Vorträge, Gäste

7.1. Tagungen

Workshop “First Workshop on Treebanks and Linguistic Theories (TLT 2002)” in Sozopol, Bulgarien, 20.–21. September 2002.

Einwöchiger Workshop “Machine Learning Approaches in Computational Linguistics” auf der 14th European Summer School in Logic, Language, and Information (ESSLLI 2002) in Trento, Italien, 5.–9. August 2002.

Herbstschule “Empirische Sprachwissenschaft und Maschinelle Sprachverarbeitung”, Sozopol, Bulgarien, 8.–22. September 2002.

Deutsch-Israelische Minerva-Herbst-Schule “Computational Linguistics” in Tübingen und Blaubeuren, Deutschland, 2.–11. Oktober 2002.

Workshop “Second Workshop on Treebanks and Linguistic Theories (TLT 2003)” in Växjö, Schweden, 14.–15. November 2003.

Einwöchiger Workshop “Combining Shallow and Deep Processing for NLP” auf der 16th European Summer School in Logic, Language, and Information (ESSLLI 2004) in Nancy, Frankreich, 9.–13. August 2004.

Workshop zum Stuttgart-Tübingen Tagset in Tübingen, 9. Dezember 2004.

Workshop “Third Workshop on Treebanks and Linguistic Theories (TLT 2004)” in Tübingen, 10.–11. Dezember 2004.

7.2. Vorträge

Hinrichs, E.W. (Jan. 2002): “Robust Syntactic Annotation of Corpora and Memory-Based Parsing”. Eingeladener Plenumsvortrag auf der *16. Pacific Asia Conference on Language, Information and Computation (PACLIC 16)*. Cheju Island, Korea.

Hinrichs, E.W. (Apr. 2002): “Morphological Disambiguation for German”. Einladungsvortrag am *Department of Linguistics, Ohio State University*, Columbus, Ohio.

Hinrichs, E.W. (Apr. 2002): “Dependency-Based Parsing for German”. Einladungsvortrag am *Linguistic Modelling Laboratory* der Bulgarischen Akademie der Wissenschaften, Sofia, Bulgarien.

Hinrichs, E.W. (März 2004): “Spoken Language Treebanks”. Einladungsvortrag auf der Winterschule *Treebanks: Formats, Tools and Usage* an der Universität Stockholm.

Hinrichs, E.W. (März 2004): Panelist auf der Podiumsdiskussion *Linguistic Theory and Natural Language Processing*, Universität Stockholm.

Hinrichs, E.W. und J. Trushkina (Sept. 2002): “Forging Agreement: Morphological Disambiguation of Noun Phrases”. TLT 2002, Sozopol, Bulgarien.

Hinrichs, E.W. und J. Trushkina (Okt. 2002): “Getting a Grip on Morphological Disambiguation”. *KONVENS 2002*, Saarbrücken.

Hinrichs, E.W. und J. Trushkina (Juli 2003): “Morphological Disambiguation of German Noun Phrases”. Einladungsvortrag auf dem *Japanese-German Workshop on Natural Language Processing*, Sapporo, Japan.

Hinrichs, E.W. und J. Trushkina (Nov. 2003): “N-gram and PCFG Models for Morpho-syntactic Tagging of German”. *TLT 2003*, Växjo, Schweden.

Hinrichs, E.W. und J. Trushkina (Jan. 2004): “Treebank Transformations for Performance Optimizations of a PCFG-based Tagger”. Konferenz des SFB 441 *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, Tübingen.

Hinrichs, E.W. und J. Trushkina (Apr. 2004): “Morpho-Syntactic Annotation of German with a Large Tagset”. Einladungsvortrag am Linguistics Department, The Ohio State University, Columbus, Ohio, USA.

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

Kübler, S. (Juni 2003): “Memory-Based Parsing for German”. Einladungsvortrag an der Bulgarischem Akademie der Wissenschaften, Sofia, Bulgarien.

Kübler, S. (Juli 2003): “Memory-Based Parsing for German”. Einladungsvortrag beim *Japanese-German Workshop on Natural Language Processing*, Sapporo, Japan.

Kübler, S. (Sept. 2003): “Parsing Without Grammar – Using Complete Trees Instead”. *RANLP 2003*, Borovets, Bulgarien.

Kübler, S. (Nov. 2003): “Parsing Without Grammar – Using Complete Trees Instead”. Einladungsvortrag an der *Alfa Informatica*, Groningen, Niederlande.

Kübler, S. (Apr. 2004): “Parsing Without Grammar – Using Complete Trees Instead”. Einladungsvortrag am Linguistics Department, The Ohio State University, Columbus, Ohio, USA.

Kübler, S. und H. Telljohann (Juni 2002): “Towards a Dependency-Based Evaluation for Partial Parsing”. LREC-Workshop *Beyond PARSEVAL – Towards Improved Evaluation Measures for Parsing Systems*, Las Palmas, Spanien.

Müller, F.H. und T. Ule (Aug. 2002): “Annotating topological fields and chunks – and revising POS tags at the same time”. *COLING 2002*, Taipei, Taiwan.

Telljohann, H., E.W. Hinrichs und S. Kübler (Mai 2004): “The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone”. *LREC 2004*, Lissabon, Portugal.

Ule, T. (Apr. 2002): “Reasons for doing linguistic annotation”. Einladungsvortrag an der Bulgarischem Akademie der Wissenschaften, Sofia, Bulgarien.

Ule, T. (Nov. 2003): “Directed Treebank Refinement for PCFG Parsing”. *TLT 2003*, Växjö, Schweden.

Ule, T. und J. Veenstra (Dez. 2003): “Making CFGs Probabilistic with Treebank Refinement”. *CLIN 2003*, Antwerpen, Belgien.

Ule, T. und K. Simov (Mai 2004): “Unexpected Productions May Well be Errors”. *LREC 2004*, Lissabon, Portugal.

Veenstra, J., F.H. Müller und T. Ule (Sept. 2002): “Topological Fields Chunking for German”. *CoNLL 2002*, Taipei, Taiwan.

Veenstra, J., F.H. Müller und T. Ule (Aug. 2002): “Topological Fields Chunking for German”. ESSLLI-Workshop *Machine Learning Approaches in Computational Linguistics*, Trento, Italien.

7.3. Projektübergreifende Aktivitäten

SFB 441, C1: Integration der TüBa-D/Z und der TüPP-D/Z in TUSNELDA.

SFB 441, A5: Recherche nach Distributionsidiosynkrasien in der TüPP-D/Z.

SFB 441, B1: Internationales Programm für Doktoranden “Empirische Sprachwissenschaft und Maschinelle Sprachverarbeitung”; Ausrichtung der Herbstschule im September 2002 in Sozopol, Bulgarien.

SFB 441, B11: Aufbereitung der Korpusdaten des Tibetischen in einem XML-strukturierten Korpus in CLARK in Kooperation mit dem Linguistic Modelling Laboratory der Bulgarischen Akademie der Wissenschaften, Sofia, Bulgarien und **C1**.

Linguistic Modelling Laboratory der Bulgarischen Akademie der Wissenschaften, Sofia, Bulgarien (K. Simov): Gastaufenthalt des Projektmitarbeiters T. Ule.

Xerox Research Centre Europe, Grenoble, Frankreich: XIP System.

Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart: Parser LoPar und elektronisches Valenzwörterbuch IMSLex.

Centrum voor Nederlandse Taal en Spraak der Universität Antwerpen, Belgien (W. Daelemans): Gastaufenthalt der Projektmitarbeiterin M. Liepert.

Arbeitsbereich Linguistik der Universität Münster (W. Paprotté): Gastaufenthalt des Projektmitarbeiters T. Ule.

7.4. Gastwissenschaftler

J. Veenstra, ILK, Universität Tilburg, Niederlande, 20.2. - 23.2.2002

D. Tufis, RACAI, Akademie der Wissenschaften, Bukarest, Rumänien, 23.11. - 27.11.2002

W. Daelemans, CNTS, Universität Antwerpen, Belgien, 3.8. - 10.8.2002 (Vortrag auf dem ESSLLI-Workshop *Machine Learning Approaches in Computational Linguistics* in Trento, Italien)

R. Mitkov und C. Orasan, Computational Linguistics and Language Engineering, University of Wolverhampton, U.K. 18.12. - 22.12.2003

M. Wolters, Queen Margaret University College, U.K., 19.4. - 23.4.2004

N.N. zwei eingeladene Gastvorträge auf der *TLT 2004*, Tübingen, 10.12. - 11.12.2004

8. Zitierte Literatur

Zu zitierten Publikationen der Projektmitglieder 2002-2004 siehe Abschnitt 6.

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

Brants, S., S. Dipper, S. Hansen, W. Lezius und G. Smith (2002): „The TIGER Treebank“, in *Proceedings of TLT 2002*, Sozopol, Bulgarien, S. 24–41.

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

Carroll, J. (Hrsg.) (2002): *Proceedings of the LREC-Workshop Beyond PARSEVAL–Towards Improved Evaluation Measures for Parsing Systems*, Las Palmas, Spanien.

Dienes, P. und C. Oravecz (2000): „Bottom-Up Tagset Design from Maximally Reduced Tagset“, in *Proceedings of LINC 2000*, Luxemburg, S. 42–47.

Dubey, A. und F. Keller (2003): „Probabilistic Parsing for German using Sister-Head Dependencies“, in *Proceedings of ACL 2003*.

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.

Duchier, D. (2000): „Configuration Of Labeled Trees Under Lexicalized Constraints And Principles“, *Journal of Language and Computation* . eingereicht Dez. 2000.

Eckle-Kohler, J. (1999): *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*, Logos, Berlin.

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*, Sapporo, Japan.

Hajic, J. und Z. Uresova (2003): „Linguistic Annotation: from Links to Cross-Layer Lexicons“, in *Proceedings of TLT 2003*, Växjö, Schweden, S. 69–80.

Höhle, T. (1986): „Der Begriff Mittelfeld, Anmerkungen über die Theorie der topologischen Felder“, in *Akten des Siebten Internationalen Germanistenkongresses 1985*, Göttingen, S. 329–340.

Klatt, S. (2002): „Combining a Rule-Based Tagger with a Statistical Tagger for Annotating German Texts“, in *Tagungsband der KONVENS 2002*, Saarbrücken.

Klein, D. und C. Manning (2003): „Accurate Unlexicalized Parsing“, in *Proceedings of ACL 2003*, Sapporo, Japan, S. 423–430.

Kudo, T. und Y. Matsumoto (2001): „Chunking with Support Vector Machines“, in *Proceedings of NAACL 2001*, Pittsburgh, PA.

Lezius, W. (2002): *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*, Dissertation, IMS, Universität Stuttgart.

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

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

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

Schiller, A. (1995): *DMOR: Benutzer-Handbuch*, Draft, IMS, Universität Stuttgart.

Schiller, A., S. Teufel und C. Thielen (1995): *Guidelines für das Tagging deutscher Textkorpora mit STTS*, Technischer Bericht, Universität Stuttgart und Universität Tübingen.

Streiter, O. (2001): „Recursive Top-Down Fuzzy Match, New Perspectives on Memory-Based Parsing“, in *Proceedings of PACLIC 2001*, Hong Kong.

Tufiş, D. (2000): „Using a Large Set of EAGLES-Compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging“, in *Proceedings of LREC 2000*, Athen, Griechenland, S. 1105–1112.

Uszkoreit, H. (1987): *Word Order and Constituent Structure in German*, Band 8 von *CSLI Lecture Notes*, CSLI, Menlo Park, CA.