Ergebnisbericht

Teilprojekt: A1

Thema: Repräsentation und Erschließung linguistischer Daten

Leiter: Erhard W. Hinrichs, Prof. Dr.

Mitarbeiter: Laura Kallmeyer, Dr., Mitarbeiterin 1/99-8/00
Sandra Kübler, Mitarbeiterin 1/99-12/00
Lothar Lemnitzer, Dr., Mitarbeiter 4/99-9/99
Frank H. Müller, assoziierter Stipendiat des GK ILS
Manfred Sailer, Dr., Mitarbeiter 10/00-12/01
Ilona Steiner, Mitarbeiterin 4/00-12/01
Heike Telljohann, Dr., Mitarbeiterin 1/01-12/01

Inhalt

1. Kenntnisstand bei der Antragstellung und Ausgangsfragestellung
2. Ergebnisse und ihre Bedeutung
3. Vergleiche mit Arbeiten außerhalb des Sonderforschungsbereichs und Reaktionen der wissenschaftlichen Öffentlichkeit auf die eigenen Arbeiten
4. Offene Fragen
5. Veröffentlichungen und Manuskripte
6. Aktivitäten: Tagungen, Vorträge, Gäste
7. Zitierte Literatur

1 Kenntnisstand bei der Antragstellung und Ausgangsfragestellung

Ziel des Projektes A1 ist es, das Verhältnis von Empirie und Theorie in der Grammatikforschung anhand von elektronisch verfügbaren Textkorpora zu untersuchen. [1]

Die Nützlichkeit elektronischer Korpora für die linguistische Theoriebildung hängt vor allem von der Art ihrer Aufbereitung ab, wobei der derzeitige Schwerpunkt auf komplexen syntaktischen Annotationen liegt. Forschungsschwerpunkte ergeben sich zum einen dadurch, daß theorieneutrale *Repräsentationsformen* verwendet werden müssen, um eine breite Skala von Anwendungsmöglichkeiten zu garantieren. Zum anderen muß die *Erschließbarkeit* der Korpora gewährleistet werden.

Diesen Zielen entsprechend wurden drei Projektbereiche (PB 1 – PB 3) definiert, die sich um die Bereitstellung komplexer syntaktischer Annotationen, die Abfrage in komplex annotierten Korpora und die weitere linguistische Aufbereitung komplex annotierter Korpora gruppieren.

Automatische inkrementelle Annotation (PB 1): Für verschiedene Annotationsebenen (vor allem Chunks und Sätze) werden Repräsentationssysteme zugrunde gelegt, die eine informationsreiche, theorieneutrale Annotation ermöglichen. Es sollten Verfahren untersucht werden, die eine inkrementelle Abbildung von einer Annotationsebene in die nächsthöhere erlauben. Zentrale Themen waren sowohl die Repräsentationssysteme als auch die Erschließung einer Annotationsebene aus der darunterliegenden durch den Einsatz (auf Chunkebene) regelbasierter und (auf Satzebene) maschineller Lernverfahren.

Annotationserschließung (PB 2): Syntaktische Annotationssysteme sind in der Regel baumbasiert. Zur systematischen Erschließung der Daten in syntaktisch annotierten Korpora wird eine Querysprache benötigt, die der Baumannotation Rechnung trägt. Hierzu sollte eine entsprechende Baumdeskripitonssprache definiert werden. In diesem Bereich ist die Deskriptionssprache als Repräsentation einer Query zentral, die dann zur Erschließung von annotierten Daten aus einem Korpus durch ein Querytool verwendet werden kann.

Theoriegeleitete Annotationsveränderung (PB 3): Es sollten Möglichkeiten untersucht werden, einen Übergang zwischen quantitativen Daten und qualitativen Daten herzustellen. Ziel war es, ein vorhandenes Korpus in ein Korpus zu überführen, das genau und nur die Information beinhaltet, die von einer bestimmten linguistischen Theorie angenommen wird. Die Datenrepräsentation betrifft hier primär den Status theorieorientierter Annotationen, die notwendigerweise immer nur partiell sind. Hinsichtlich der Datenerschließung wird die Abbildung von robusten, vollständigen Annotationen zu detaillierten, partiellen untersucht.

[1] In diesem Kapitel wird lediglich die Literatur angemerkt, die im Erstantrag A1 nicht erwähnt wurde.

1.1 Automatische inkrementelle Annotation

Dieser Projektbereich befaßte sich mit der zentralen Fragestellung, wie Korpora automatisch annotiert werden können. Da die automatische syntaktische Annotation ein sehr komplexer Prozeß ist, bei dem sich vor allem bei der Behandlung von Anbindungsambiguitäten und bei rekursiven Strukturen exponentielles Verhalten sowohl bei Fehlerquoten als auch bei den Parsezeiten ergeben, sollte hier ein inkrementeller Ansatz verfolgt werden. Als erster Schritt sollte ein Part-of-Speech-Tagging unter Verwendung des STTS durchgeführt werden, danach ein partielles Parsen mittels eines Chunk-Parsers, da ein Chunk-Parser robust und effizient die Annotation auf der Phrasenstruktur durchführen kann. Als letzter Schritt sollte eine komplette Annotation mit topologischen Feldern und Funktor-Argument-Strukturen angegangen werden.

Chunks sind von ihrer Konzeption her als nichtrekursiv definiert worden. Jedoch treten im Deutschen im Gegensatz zum Englischen Strukturen auf, in denen bei einer nichtrekursiven Handhabung von Chunks 'gestrandete' Konstituenten auftreten (vgl. (1)). Diese Problemstellung und die Frage, ob auch komplexe Chunk-Strukturen wie in (2) flach annotiert bleiben sollten, waren Gegenstand der Überlegungen hinsichtlich der Übertragung des Chunk-Begriffes auf das Deutsche.

(1) a. * the proud [PP of his son] man
    b. der [PP auf seinen Sohn] stolze Mann

(2) [PP bei einer kurzen, technisch bedingten Zwischenlandung]

Weiterhin war zu untersuchen, wie das Chunking-System als ein Modul in ein komplexes, mehrstufiges Parsing-System eingebaut werden kann. So sollte die Wartung vereinfacht werden und der Ablauf der inkrementellen Annotation, bei der eine Ebene der Annotation auf der vorhergehenden aufbaut, besser untersucht werden können. Die Teilung der Annotations-Ebenen in Tagging, Chunking und Satzannotation warf ebenfalls die Frage nach der Definition der Chunks im Kontext der vorhergehenden und der weiteren Annotation auf, sowie die Frage nach der Kommunikation der unterschiedlichen Ebenen. Hierbei stand ein Ausgleich zwischen linguistischer Anforderung an die Chunk-Analyse und der Komplexität der Anwendung im Vordergrund, um die empirische Verwendbarkeit von Chunks zu gewährleisten.

Um für die linguistische Theoriebildung relevant zu sein, müssen die Annotationen bestimmte komplexe syntaktische Information wie z.B. die Funktor-Argument-Struktur eines Satzes abbilden. Die Chunk-Strukturen bieten sich für die automatische Annotation komplexer Syntaxbäume als Grundlage an, da sie die grundlegende Baumstruktur auf Phrasenebene modellieren und somit eine solide Ausgangsannotation für die weitere Annotation bilden.

Als Parsing-Ansatz im letzten Schritt sollte ein *memory-based*-Verfahren angewandt werden, da für dieses Verfahren gute Ergebnisse bei nicht-lokalen Beziehungen vorliegen. Dieser Ansatz ermöglicht weiterhin einen Zugriff auf lexikalische Eigenschaften, was vor allem bei dem Problem der PP-Anbindung nötig ist. Trainingsdaten für den Parser waren in Form der *Tübinger Baumbank des gesprochenen Deutschen* (TüBa-D) (Stegmann, Telljohann und Hinrichs 2000) vorhanden, die im Rahmen von Verbmobil in Tübingen annotiert wurde.

1.2 Annotationserschließung

Im Zentrum des zweiten Projektbereichs stand die Aufgabe, eine intelligente Anfragesprache für syntaktisch annotierte Korpora zu entwickeln. Hierbei war zunächst zu klären, welche in syntaktischen Annotationen enthaltene Information linguistisch relevant ist, mit anderen Worten, welche Anfragen an ein syntaktisch annotiertes Korpus idealerweise möglich sein sollten. Anschließend sollte, basierend auf einer geeigneten Baumlogik, eine Anfragesprache entwickelt werden. Für die gewählte Baumbeschreibungssprache sollte eine denotationelle Semantik entwickelt werden, so daß eine direkte modell-theoretische Verbindung zwischen einem Ausdruck in der Anfragesprache und den gesuchten sytaktischen Strukturen besteht.

Diese Baumbeschreibungssprache sollte unabhängig sein von Überlegungen, die den Suchvorgang selber betreffen. Das heißt, diese Sprache stellt gegenüber bestehenden Anfragesprachen eine weitere Abstraktionsstufe dar. Hierbei sind formale Eigenschaften der Baumbeschreibungssprache von großem Interesse, insbesondere Resultate zur Entscheidbarkeit und Komplexität. Es mußte abgewogen werden zwischen der aus linguistischen Gründen gewünschten Ausdrucksstärke auf der einen Seite und der Komplexität der gewählten Logik auf der anderen Seite. Diese Fragen sollten im Rahmen des Projektes untersucht werden.

Es ist von großem Interesse, nicht nur auf der Ebene der Ketten von Wörtern sondern auch bezogen auf die syntaktische Struktur Information aus einem Korpus zu extrahieren. Das heißt, daß es eine geeignete Anfragesprache erlauben sollte, nach bestimmten Teilstrukturen zu suchen und darüber hinaus sogar einzelne Eigenschaften dieser Teilstrukturen offen bzw. unterspezifiziert zu lassen (z.B. die Unterspezifikation der linearen Abfolge).

Ist eine adäquate Baumbeschreibungssprache entwickelt, so sollte unter Verwendung dieser Sprache zunächst exemplarisch für die Baumbank TüBa-D der Prototyp eines Query-Systems implementiert werden. Dieses System sollte den anderen Projekten des SFB 441 zur Verfügung gestellt werden.

1.3 Theoriegeleitete Annotationsveränderung

Ziel des dritten Projektbereichs war es, den Brückenschlag zwischen quantitativen und qualitativen Daten auf der Ebene syntaktisch annotierter Korpora zu schaffen.

Syntaktisch annotierte Korpora folgen in der Regel einem *theorieneutralen*, *robusten*, aber auch relativ *groben* Annotationsschema. Linguistische Analysen sind hingegen stets *theoriegeleitet*, in der Regel nur *partiell*, d.h., auf wenige Phänomene beschränkt, und dafür sehr *fein* hinsichtlich der Analyse dieser Phänomene. Aus diesem Gegensatz ergibt sich, daß Baumbanken ungeeignet sind, linguistische Einzelanalysen an natürlichen Daten zu illustrieren.

Zum Zeitpunkt der Antragstellung gab es zwei Möglichkeiten, syntaktisch annotierte Korpora phänomenorientiert aufzubereiten: (i) phänomenorientiertes Tagging, oder (ii) Parsen mittels *Fragmentgrammatiken*. Im ersten Fall wird das Annotationsschema leicht erweitert, um seine Grobheit zumindest hinsichtlich eines bestimmten linguistischen Phänomenbereichs zu reduzieren. Dieser Ansatz wurde im Teilprojekt C1 bei der Erstellung des TUSNELDA-Korpus verfolgt. Das Erstellen einer Fragmentgrammatik setzt drei Dinge voraus: eine vollständige linguistische Analyse aller Daten, einen hohen Grad an Formalisiertheit des theoretischen Ansatzes und entsprechende automatische Verarbeitungswerkzeuge (Parser). Keine dieser drei Vorbedingungen kann als uneingeschränkt für alle linguistischen Theorien gegeben angesehen werden.

Ausgehend von eigenen Vorarbeiten zur Formalisierung der *Head-Driven Phrase Structure Grammar* (HPSG, Pollard und Sag (1994)) und der ontologischen Konzeptualisierung von Baumbanken als *Modell* eines Annotationsschemas in PB 2, sollte der Frage nachgegangen werden, welche Form eine Sammlung theoriegeleiteter Annotationen sinnvollerweise annehmen kann. Hierbei stellten sich uns drei Alternativen dar, die ihre Widerspiegelung in der HPSG-Formalisierung haben:

(i) Die theoriegeleitete Annotation ist ein Modell, ebenso wie das theorieneutrale Annotationsschema. Dabei sind die Annotationen vollständig, ebenso wie das für Objekte in der Denotation einer HPSG-Grammatik im Sinne von Pollard und Sag (1994) gilt. (ii) Die Annotation ist zwar ein Modell, allerdings läßt das Modell *partielle Information* zu, d.h., die Objekte im Modell sind nicht notwendigerweise vollständig. Dies entspricht dem Geist von HPSG-Grammatiken von Pollard und Sag (1987). (iii) Die theoriegeleitete Annotation ist eine *Beschreibung*, kein Modell. Diese Sichtweise ist ebenso wie die erste im Einklang mit Pollard und Sag (1994), allerdings enthalten die Annotationen, ebenso wie unter (ii) lediglich die für die konkrete Einzeluntersuchung relevante Information. Aus dieser Einschätzung war bereits zum Zeitpunkt der Antragsstellung eine Präferenz zugunsten der letztgenannten Alternative deutlich.

Zu Beginn der Förderungsphase waren mit dem Korpus TüBa-D und dem dazugehörigen Annotationsschema die technischen Vorraussetzungen für die zu untersuchende Fragestellung gegeben. Durch die tiefgehenden Untersuchungen zum Verbalkomplex und zur Verbstellung im Deutschen, wie sie an der Universität Tübingen bereits in vorangegangenen Projekten im SFB 340 vorlagen, und wie sie weiterführend auch in Partnerprojekten des SFB 441 vorgesehen waren, konnte ebenfalls eine solide Forschungsbasis für eine theoriegeleitete Annotation in diesem Phänomenbereich vorausgesetzt werden. Schließlich hat die bereits Ende der 90er Jahre intensiv geführte

Diskussion über die Beziehung zwischen Modell und Beschreibung in Grammatiktheorien zu einer pointierten Ausrichtung der Grundfragestellung dieses Projektbereichs beigetragen.

2 Ergebnisse und ihre Bedeutung

Für die drei Projektbereiche (PB 1 – PB 3) lassen sich folgende Ergebnisse zusammenfassen. In PB 1 wurde eine hybride Parsingarchitektur zur komplexen syntaktischen Annotation von Textkorpora entwickelt, die *finite-state*-Methoden mit maschinellen Lernverfahren des *memory-based learning* verbindet. Das in PB 2 entwickelte Query-Tool VIQTORYA ermöglicht es, in diesen syntaktisch annotierten Korpora nach komplexen syntaktischen Strukturen zu suchen. Die Überführung dieser theorieneutralen syntaktischen Repräsentation in die Spezifikation linguistischer Einzeltheorien wurde in PB 3 anhand des Verbalkomplex im Deutschen untersucht. Eine ausführliche Darstellung folgt.

2.1 Automatische inkrementelle Annotation

Die Annotation auf POS-Ebene wurde mit bereits verfügbaren Programmen ausgeführt (LIKELY (Feldweg 1995), TnT (Brants 2000)). Zur manuellen Erstellung von Trainingsdaten wurde eine Studie in Kooperation mit dem Projekt C1 durchgeführt und die Ergebnisse in Kübler und Wagner (2000) vorgestellt.

Es wurde ein Chunking-System entwickelt, das auf der Grundlage der bloßen Wortarteninformation in der Lage ist, ein aus verschiedenen Textsorten bestehendes Korpus mit komplexen und heterogenen Chunk-Strukturen robust zu annotieren.

Da sich Chunk-Strukturen in den untersuchten Texten als weitaus komplexer darstellten als im Englischen und als sie sich in bisher im Deutschen analysierten Texten dargestellt hatten, wurde entschieden, Chunks nicht flach zu annotieren, sondern Chunks innerhalb von Chunks zuzulassen. Da jedoch auch weiterhin keine Rekursivität zugelassen werden sollte, um Anbindungsprobleme auf dieser Ebene zu vermeiden, wurden Chunks so definiert, daß sie nichtrekursive, ununterbrochene Kernphrasen umfassen. Das bedeutet für die Beispiele (3) und (4), daß der Nominal-Chunk (NC) *stolze Mann* nicht den Präpositional-Chunk (PC) *auf seinen Sohn* enthalten kann. Eine Annotation solcher Strukturen wäre mit der benutzten Methode der kaskadierten *finite-state*-Transduktoren nicht möglich, ohne deren Robustheit und Schnelligkeit zu beeinträchtigen. Die obige Struktur bleibt somit nur partiell annotiert und der Artikel *der* gestrandet (vgl. (3)).

(3) der [$_{PC}$ auf [$_{NC}$ seinen Sohn]] [$_{NC}$ [$_{AJAC}$ stolze] Mann]

(4) [$_{PC}$ bei [$_{NC}$ einer [$_{AJACC}$ [$_{AJAC}$ kurzen], [$_{AJAC}$ technisch bedingten]] Zwischenlandung]]

Die Aufteilung in zwei Annotationsaufgaben hat dabei nicht allein methodische Gründe. Sie spiegelt auch wider, daß die beiden Aufgaben sich auf unterschiedliche Ebenen der linguistischen Beschreibung beziehen. Denn, während die Chunk-Annotation auf syntaktische Restriktionen zurückgreift, benötigt die Ebene der weiteren Annotation lexikalische Selektionsbeschränkungen. Eben diese Unterscheidung markiert die Trennung zwischen flachem Parsen und globalem Parsen. Dies bedeutet beispielsweise für die Behandlung von Adverbien in Nominal-Chunks, daß sie nur dann als zum Chunk gehörig annotiert werden, wenn syntaktische Restriktionen dies eindeutig zeigen. So wird der Nominal-Chunk eingegrenzt auf der einen Seite durch Wörter, die eindeutig auf den Anfang eines Nominal-Chunks hinweisen, und auf der anderen Seite durch das Kopf-Nomen. Adverbien, die dazwischen liegen, werden zum Nominal-Chunk gezählt. Adverbien, die nicht dazwischen liegen, werden aufgrund ihrer unklaren Anbindung nicht dazugezählt. So ist die Anbindung in (5) mehrdeutig, da *so* auch ein Konjunktionaladverb sein kann, während die Anbindung in (6) eindeutig ist, da die Präposition *mit* den unmittelbar folgenden Beginn eines NC markiert. Auf lexikalische Informationen zur Bestimmung der Anbindung von Token wird nur in wenigen, im wesentlichen auf die Funktionswörter beschränkten Fällen zurückgegriffen. Auf morphologische Informationen mußte nur für das Tag-Verbesserungssystem zurückgegriffen werden.

(5) Ich habe so [$_{NC}$ viele Leute] getroffen.

(6) Ich habe mich mit [$_{NC}$ so vielen Leuten] getroffen.

Eine besondere Funktion bei der Chunk-Analyse kam den Verb-Chunks zu, da sie die Schnittstelle zwischen Chunks und topologischen Feldern bilden. So sind die Verb-Chunks Teile der Satzklammer, die für die Erkennung von topologische Feldern unerlässlich ist. Der Erkennung der unterschiedlichen Typen von Verb-Chunks, die viel aussagt über die weitere Struktur des Satzes, wurde daher besondere Aufmerksamkeit gewidmet (vgl. Müller (2001)). Da gerade bei den Verbal-Chunks die Fehlerquote der Tags überdurchschnittlich hoch war (Faktor 1,75) und gerade die falsche Erkennung dieser Chunks die Anzahl von Folgefehlern potenziert, wurde ein Tag-Verbesserungssystem entwickelt, das die Fehlerrate hier deutlich reduzierte (um 30–50%) und somit zur Absicherung der Annotationsqualität beitrug (vgl. Müller und Ule (2001)).

Bei der Erweiterung der Chunkstrukturen zu vollständigen Bäumen lag die Hauptaufgabe in der Entwicklung geeigneter Parsingstrategien, die zum einen die von der Chunkanalyse annotierten Phrasenstrukturen als Ausgangsbasis für die weitere Annotation verwenden und die zum anderen ähnlichkeitsbasiert die beste Baumstruktur für einen Eingabesatz ermitteln können. Um zweiteres zu erreichen, mußte die Aufgabe als eine Klassifikationsaufgabe definiert werden.

	Minimum	Maximum	Durchschnitt
Precision	76.82%	77.87%	77.23%
Recall	66.90%	67.65%	67.28%
Crossing Accuracy	93.44%	93.95%	93.70%

Tabelle 1: Quantitative Evaluierung

Die hier gewählte Strategie sieht vor, daß die Klassen aus kompletten Bäumen bestehen, daß also für jeden Eingabesatz der ähnlichste Baum aus der Instanzenbasis ermittelt wird. Eine wichtige Teilaufgabe bei dieser Methode ist die Definition der Ähnlichkeitsfunktion. Bei bisherigen Fragestellungen, die mit memory-based Ansätzen klassifiziert wurden, wurden die Merkmale aus einem sehr lokalen Kontext genommen. Bei einer satzweisen Klassifikation kann von dieser Vereinfachung jedoch nicht mehr ausgegangen werden. Die Klassifikation muß bei dieser Aufgabenstellung auch die interne Struktur, d.h. vor allem die sequentielle Information der zu klassifizierenden Einheit, berücksichtigen. Aus diesem Grund mußte eine besondere Suchroutine definiert werden, die diesen Anforderungen gerecht wird (für Details siehe Kübler und Hinrichs (2000) und Kübler und Hinrichs (2001*b*)).

Damit auch die Generalisierungsfähigkeit des Verfahrens erhalten bleibt, mußte desweiteren sichergestellt werden, daß auch bisher nicht gesehene Bäume richtig annotiert werden, soweit dies möglich ist. Aus diesem Grund wurde ein zweiter Schritt implementiert, der den gefundenen ähnlichsten Baum an die Eingabe anpaßt. Da die automatische Annotation vor allem als Unterstützung für die Annotation von Baumbanken konzipiert ist, wurde bewußt darauf verzichtet, Bäume mit einer geringen Ähnlichkeit zum Eingabesatz zu akzeptieren. Stattdessen wurde ein *backing-off* Ansatz realisiert, der auf eine maximal informative Annotation auf Chunk-Ebene ausweicht.

Als Ergebnis dieser Strategien entstand ein sehr effizientes Parsing-System (vgl. auch Kübler und Hinrichs (2000), Kübler und Hinrichs (2001*a*), Kübler und Hinrichs (2001*b*)), das anhand der Baumbank TüBa-D trainiert und getestet wurde. Die Resultate der Evaluierung sind in Tabelle 1 dargestellt, zur Problematik der Evaluierung partieller Parser vgl. Kübler und Hinrichs (2001*b*). Die Systemausgabe erfolgt in XML und kann als Eingabe für das Query-Tool aus PB 2 dienen. Das System kann bei Bedarf den anderen Projekten, die sich mit deutschsprachigen Korpora befassen, zur Verfügung gestellt werden.

2.2 Annotationserschließung

Um nach linguistisch interessanten Phänomenen in einem Korpus suchen zu können, sind zwei Voraussetzungen notwendig: ein Korpus, das reich genug annotiert ist, um die relevanten Eigenschaften dieser Phänomene zu kodieren, sowie ein Query-Tool mit einer Anfragesprache, die es erlaubt, die jeweiligen Eigenschaften auszudrücken.

Korpora, die für syntaktische Fragestellungen die erste Voraussetzung erfüllen, sind z.B. die TüBa-Baumbanken. Die entwickelte Anfragesprache und der Prototyp eines adäquaten Query-Tools werden in 2.2.1 bzw. 2.2.2 vorgestellt.

Bevor eine intelligente Anfragesprache entwickelt werden konnte, war zunächst zu klären, welche strukturellen Eigenschaften linguistisch interessant sind und welche Konsequenzen sich für die zu wählende Baumlogik (Anfragesprache) ergeben. Vorarbeiten zu Baumbeschreibungen, allerdings im Rahmen von Baumgrammatiken, finden sich in den Arbeiten Kallmeyer (1999*a*), Kallmeyer (1999*b*), Kallmeyer (1999*c*), Kallmeyer (2000*d*), Kallmeyer (2001). Um die nötige Ausdrucksstärke der Anfragesprache genauer zu bestimmen, wurde in den Projekten des SFB eine Bedarfsprüfung durchgeführt und weiterhin die auch in PB 1 betrachtete Baumbank TüBa-D zugrundegelegt. Bei der Entwicklung der Baumbeschreibungssprache spielten auf der einen Seite formale Eigenschaften der zugrundeliegenden Logik eine wichtige Rolle, auf der anderen Seite sollte die Anfragespache aber auch ausdrucksstark genug sein, um nach linguistisch relevanten Konstruktionen suchen zu können. Die resultierende Anfragesprache ist in 2.2.1 beschrieben.

Ein Beispiel für eine linguistisch relevante Konstruktion ist die NP-PP-Aufspaltung, die z.B. in De Kuthy (2000) und in Hinrichs und Nakazawa (1999) untersucht wurde. Eine derartige Konstruktion, annotiert nach dem Annotationsschema der TüBa-D, ist in Abbildung 1 dargestellt. Um nach solchen Konstruktionen in einem deutschen Korpus suchen zu können, muß es z.B. möglich sein, nach einer Präpositionalphrase (*über Chomsky*) zu suchen, die im Vorfeld steht und das Akkusativobjekt modifiziert, sowie nach einem Akkusativobjekt (*ein Buch*), welches im Mittelfeld steht.

Zur Frage der linguistischen Ausdrucksstärke der Anfragesprache ergaben sich fruchtbare Diskussionen mit den Projekten A3 und B3 über die linguistische Relevanz bestimmter Phänomene und die Reichhaltigkeit der Annotationen in existierenden Korpora. Die Komplexität der entsprechenden Anfragen, spezielle Anfragen zur Extraktion von langen Abhängigkeiten, ist im Rahmen des Projekts untersucht worden (siehe Kallmeyer (2000*a*) und Kallmeyer (2000*b*)).

2.2.1 Die Anfragesprache

Die in Kallmeyer (2000*a*) und Kallmeyer (2000*b*) durchgeführten theoretischen Untersuchungen ergaben, daß selbst für relativ komplexe strukturelle Phänomene Anfragen in einer Logik erster Ordnung mit nur wenigen Variablen und ohne Negation und Quantoren ausreichen. Dies führte zu der Entscheidung, sich bei der Wahl der Anfragesprache auf ein Fragment einer Logik erster Ordnung zu beschränken: Bei der entwickelten Anfragesprache handelt es sich um eine Prädikatenlogik erster Ordnung ohne Quantoren, wobei die Variablen existentiell interpretiert werden. Negation ist nur für atomare Formeln zugelassen.

Die Anfragsprache erlaubt es, in den Bäumen eines Korpus nach Wörtern zu suchen,

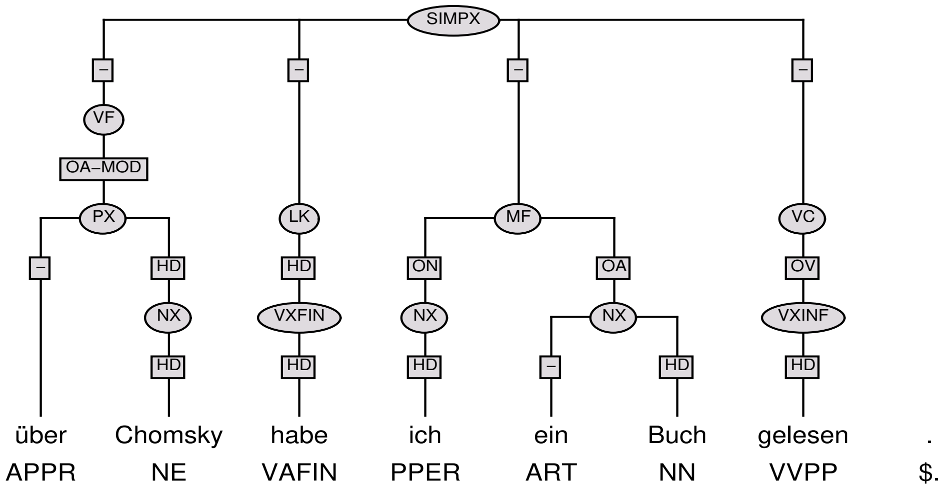

Abbildung 1: NP-PP-Aufspaltung

nach Wortarten, syntaktischen Kategorien, grammatische Funktionen und topologischen Feldern sowie nach den Relationen: *Dominanz*, *unmittelbare Dominanz* und *lineare Präzedenz* zwischen Knoten. Alle genannten Prädikate und Relationen können außerdem negiert werden. Komplexe Anfragen können mit Hilfe der logischen Operatoren Konjunktion und Disjunktion formuliert werden. Somit ist es möglich, Teilstrukturen zu definieren, wobei außerdem bestimmte Eigenschaften dieser Teilstrukturen (Dominanzbeziehung bzw. die lineare Abfolge von Knoten) unterspezifiziert werden können. Diese Möglichkeiten der Unterspezifikation sind z.B. wichtig für die Extraktion von diskontinuierlichen Konstituenten wie die bereits erwähnte NP-PP-Aufspaltung. Eine mögliche Anfrage zur Extraktion derartiger NP-PP-Aufspaltungen findet sich unter “Query” in Abb. 3. Hier wird nach Strukturen gesucht, die eine das Akkusativobjekt modifizierende Präpositionalphrase (Variable 1) enthalten, die vom Vorfeldknoten (Var. 2) dominiert wird (>> = Dominanz), und die außerdem im Mittelfeld eine Nominalphrase (Variable 3) als Akkusativobjekt enthalten.

Eine Definition der Anfragesprache und der zugrundeliegenden modelltheoretischen Semantik findet sich in Kallmeyer (2000*c*). Der Verzicht auf Quantoren und die Einschränkung auf atomare Negationen bedeuten eine Verringerung der Komplexität der Logik und tragen wesentlich dazu bei, daß eine effiziente Suche in einem Korpus möglich ist. Bei den für die Logik definierten intendierten Modellen handelt es sich nicht eigentlich um Bäume, sondern um baumähnliche endliche Graphen. Im Gegensatz zu Bäumen besitzen diese Graphen nicht immer eine eindeutige Wurzel und zwei Knoten innerhalb eines Graphen haben nicht notwendig eine Dominanz- oder lineare Präzedenzbeziehung, wie es bei Bäumen der Fall ist. Diese Überlegungen zu den der Annotation und damit auch der Anfragesprache zugrundeliegenden linguistischen

Datenstrukturen wurden intensiv mit Teilprojekt A2 diskutiert.

2.2.2 Das Query-Tool VIQTORYA

Das Query-Tool VIQTORYA (A **Vi**sual **Q**uery **T**ool f**or** **Y**ntactically **A**nnotated Corpora) ist ein Prototyp für syntaktisch annotierte Korpora und verwendet die entwickelte Anfragesprache.[2] VIQTORYA erlaubt es, in Korpora zu suchen, die auch von Bäumen abweichende Datenstrukturen enthalten, genauer baumähnliche Strukturen, die unangebundene Teilbäume enthalten können, wie es in den TüBa-Baumbanken der Fall ist.

Die grundlegende Idee des Query-Tools besteht darin, die für die Suche relevanten Informationen in einer relationalen Datenbank abzulegen und die Anfragen auf der Datenbank auszuwerten. Das Tool besteht aus:

- einer Initialisierungskomponente[3], die die relevante Information (z.B. Knoten/Knotenpaare) aus den Baumstrukturen extrahiert und diese in der Datenbank ablegt. Dies muß für jedes Korpus nur einmal durchgeführt werden.
- einer graphischen Benutzerschnittstelle, um die Anfragen benutzerfreundlich spezifizieren zu können, die dann an die Anfragekomponente weitergegeben werden.
- einer Anfragekomponente, die einen Ausdruck in der oben vorgestellten Anfragesprache in eine SQL-Anfrage übersetzt und diese dann auf der Datenbank auswertet.
- Die Suchergebnisse können im Annotierungswerkzeug *Annotate* (Brants und Skut 1998) dargestellt werden.

Das Tool wurde in Java implementiert und benutzt JDBC (Java Database Connectivity) als Interface und MySQL als Datenbank-Management-System.

Ein Vorteil dieser Architektur ist, daß ein beträchtlicher Teil der Arbeit vom Datenbank-Management-System übernommen wird und nicht implementiert werden muß. Außerdem können die Indexierungsmöglichkeiten in MySQL benutzt werden, um direkt die Performanz der Suche zu verbessern.

Das relationale Datenbankschema Die Baumbank TüBa-D besteht aus mehreren Subkorpora. Für jedes Subkorpus *i* existieren in der relationalen Datenbank Tabellen für Knoten (tokens_*i*) und Knotenpaare (node_pair_*i*). Weiterhin enthält die Datenbank zwei globale Tabellen: eine Tabelle für Klassen von Knoten (node_class)

[2] In der aktuellen Prototyp-Version sind alle Elemente der Anfragesprache implementiert, mit Ausnahme der Disjunktion.

[3] Die Initialisierungskomponente wurde zunächst exemplarisch für die Baumbank TüBa-D entwickelt, sie kann aber prinzipiell auch für andere Korpora erweitert werden.

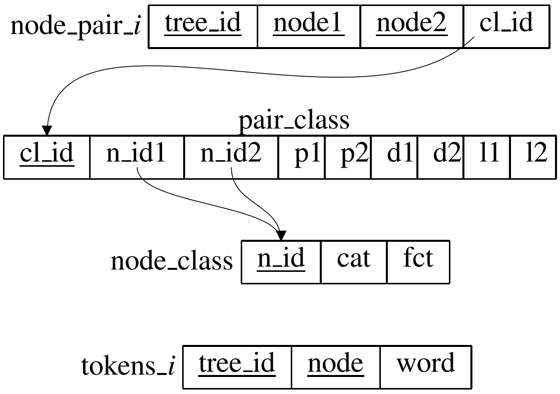

Abbildung 2: Das relationale Datenbankschema in VIQTORYA

in Abhängigkeit der syntaktischen Kategorie (Knotenlabel) und der grammatischen Funktion (Kantenlabel). Eine weitere Tabelle (pair_class) enthält Klassen von Knotenpaaren in Abhängigkeit der beiden Knotenklassen (n_id1 und n_id2) und der Relationen zwischen den beiden Knoten. Das Datenbankschema ist in Abbildung 2 dargestellt. Dieses Design ermöglicht einen effizienten Zugriff auf die durch die Anfragesprache definierten Teilstrukturen im Korpus. Genauere Ausführungen zur Architektur und zum zugrundeliegenden Datenbankschema von VIQTORYA sind in Kallmeyer (2000*c*) zu finden.

Die Benutzerschnittstelle Für die Formulierung der Suchanfragen wurde eine graphische Benutzeroberfläche entwickelt[4]. Der Anwender kann auf benutzerfreundliche Weise ein zu durchsuchendes Subkorpus anwählen, Knoten mit ihren Eigenschaften und Relationen zwischen Knoten spezifizieren. Dabei muß der Anwender die Anfragsprache nicht beherrschen. Die Anfrage wird automatisch generiert. Alternativ kann der Ausdruck natürlich auch in der Anfragesprache eingegeben werden. Weiterhin können die spezifizierten Anfragen vom Benutzer abgespeichert und für spätere Anwendungen wiederverwendet oder verfeinert werden. Die Oberfläche ist in Abbildung 3 dargestellt und enthält eine Anfrage für die bereits erwähnte NP-PP-Aufspaltung. Die Suchergebnisse können im *Annotate*-Tool dargestellt werden, ein Annotierungswerkzeug, das innerhalb des SFB zunehmend verwendet wird.

Das Query-Tool kann nun mit den Ergebnissen aus PB 1 kombiniert werden, so daß sich ein System ergibt, das es erlaubt, ein gegebenes Korpus zunächst automatisch syntaktisch zu annotieren und anschließend Anfragen an dieses Korpus zu stellen. Details des Datenbankschemas wurde mit den Projekten A3 und B10 diskutiert und das Tool wurde allen Projekten des SFB zur Verfügung gestellt.

[4]Implementiert von Thomas Schurtz.

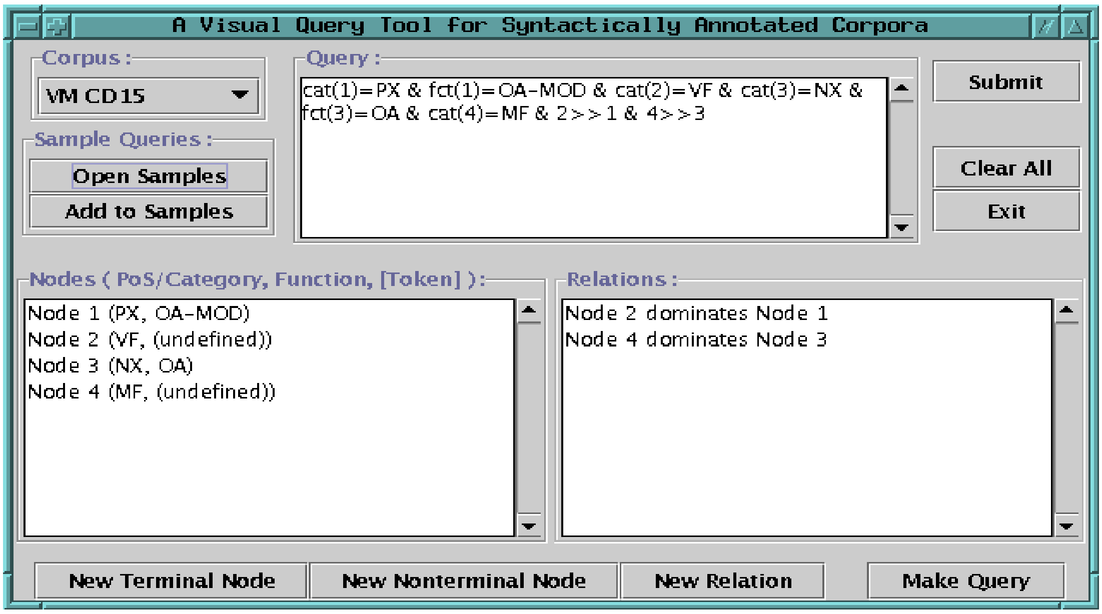

Abbildung 3: VIQTORYA's graphische Benutzeroberfläche

2.3 Theoriegeleitete Annotationsveränderung

Die Ergebnisse des PB 3 lassen sich in folgende drei Gruppen gliedern: (1) exemplarische Spezifikation einer empirischen Domäne für die theoriegeleitete Datenerschließung; (2) Exploration von Repräsentationsformen/-arten und von Abbildungsmöglichkeiten zwischen ihnen; und (3) Instantiierung einer Abbildung zur Überführung von theorieneutralen Strukturen (PB 1) in theoriegeleitete Strukturskizzen.

Durch die rasche Verbreitung neuerer Standards der Texttechnologie, vor allem im Umfeld von XML, und deren zunehmende Anwendung im linguistischen und insbesondere korpuslinguistischen Bereich wurde die Realisierung der Ziele des PB 3 wesentlich vorangetrieben.

Empirische Domäne: Mit der Dissertation von Detmar Meurers (Meurers 1999) stand eine umfassende, empirisch fundierte und formal präzise Arbeit zu den Stellungsvarietäten finiter und infiniter Verben im Deutschen zur Verfügung, die sich als Standard für die Arbeiten dieses Projektbereichs anbot. Diese Arbeit entstand im fruchtbaren Austausch der Erfahrungen des SFB 340 mit den im SFB 441 untersuchten Korpora und entwickelten Werkzeugen zur Korpusrecherche (Meurers 2000). Das Besondere an dieser Arbeit ist die Verbindung von lexikalischen Eigenschaften einzelner Verb(klass)en und allgemeinen phrasenstrukturellen Kombinationsmöglichkeiten. Basierend auf Korpus- und Literaturrecherchen wurden diejenigen Verben des Deutschen, die in Verbalkomplexen auftreten können (Tempusauxiliare, Modalverben, AcI-

Verben, etc.) hinsichtlich ihrer Stellungs-, Statusrektions- und Subkategorisierungseigenschaften klassifiziert. Durch eine genaue Prüfung der in der HPSG gegebenen Selektionsmechanismen (Kopf-Komplement, Adjunkt-Kopf, Marker-Kopf) konnten Stellungsbesonderheiten, die mit Statusirregularitäten einhergehen, als Instanzen von Marker-Kopf-Selektion identifiziert werden.

Repräsentationsformen/-arten und Abbildungsformalismus: In der Förderungsphase wurden wichtige formale Ergebnisse hinsichtlich des Status von Repräsentationen erzielt. Thematisch verwandte Fragen wurden auch im Projekt A2 angegangen, so daß es hier zu einer fruchtbaren Zusammenarbeit kam.

In Richter, Sailer und Penn (1999) wurde, aufbauend auf den Arbeiten von Paul King, eine den Anforderungen der HPSG entsprechenden linguistischen Deskriptionssprache, RSRL (*Relational Speciate Re-entrant Language*) definiert. Für sie hat Richter (2000) eine modelltheoretische Deutung gegeben und aus verschiedenen ontologischen Perspektiven präsentiert. Dabei wurden entweder linguistische Objekte selbst (*tokens*, King (1999)), Abstraktionen darüber (*types*, Pollard (1999)) oder der in ihnen enthaltene Informationsgehalt (Pollard und Sag 1987) als Elemente des linguistisch intendierten Modells betrachtet. Somit wurden der oben abgesteckte Alternativenraum für theoriegeleitete Annotationen aus formal-theoretischer Sicht durchleuchtet.

Diese Arbeiten fanden auch Eingang in die grammatiktheoretischen Untersuchungen in A2. So wird in Kepser (2000) eine coalgebraische Modellkonstruktion für die HPSG entworfen und in Kepser (2001) Fragen der Komplexität von RSRL untersucht.

Als eine Fallstudie im Übergang von Objekt und Beschreibung wurde in Sailer (2000) eine RSRL-Grammatik für Terme einer semantischen Deskriptionssprache, Ty2 (Gallin 1975, Zimmermann 1989), diskutiert. Aus der Sicht der Semantik sind solche Terme Beschreibungen, die semantische Objekte eines bestimmten semantischen Typs denotieren. In einem RSRL-Modell treten diese Terme jedoch als Objekte auf. Solche Objekte sind von linguistischer Relevanz als *logische Formen* sprachlicher Zeichen. Durch diese Arbeit wurde exemplarisch gezeigt, daß der Übergang von Beschreibung zu Objekt prinzipiell möglich ist, die genaue ontologische Bestimmung von der konkreten Zielsetzung abhängt und weitreichende Konsequenzen für die Kodierung hat.

In den genannten Arbeiten wurden die theoretischen Grundlagen für die Arbeit in PB 3 bereitgestellt und vertieft und der Ausgangsfragestellung an formal gut untersuchten Beispielen nachgegangen.

Texttechnologischer Einfluß: Das Ziel der Umannotation stellte sich für das Projekt in folgender Weise: Als Eingabestrukturen wurden Sätze genommen, die gemäß des TüBa-D-Annotationsstandards (Stegmann et al. 2000) annotiert sind. Als Ausgabestrukturen werden AVM-Beschreibungen anvisiert, die der syntaktischen Analyse der im Projekt B8/SFB 340 implementierten Fragmentgrammatik entsprechen. Allerdings sollten im Gegensatz zur Fragmentgrammatik einzig und allein diejenigen Aspekte der Zielstrukturen beschrieben werden, die für den Phänomenbereich Verbbewegung/Verbalkomplex relevant sind. Um eine Überprüfbarkeit der Ergebnisse zu

```
<s id="t84">
<tn cat="simpx">
  <f cat="vf" edge="-">
    <n cat="nx" edge="on">
      <w cat="ne" edge="hd">sarah</w>
    </n>
  </f>
  <f cat="lk" edge="-">
    <n cat="vxfin" edge="hd">
      <w cat="vvfin" edge="hd">liebt</w>
    </n>
  </f>
  <f cat="mf" edge="-">
    <n cat="nx" edge="oa">
      <n cat="nx" edge="hd">
        <w cat="art" edge="-">ein</w>
        <w cat="ne" edge="hd">foto</w>
      </n>
      <n cat="px" edge="-">
        <w cat="appr" edge="-">von</w>
        <n cat="nx" edge="hd">
          <w cat="ne" edge="hd">Peter</w>
        </n>
      </n>
    </n>
  </f>
</tn>
</s>
```

Abbildung 4: XML-Kodierung für Satz (7)

ermöglichen, wurden nur solche Sätze überführt, denen auch innerhalb der Fragmentgrammatik eine vollständige Analyse zugewiesen werden kann.

Für die Abbildung und die Definition der Zielstrukturen wurde auf neuere texttechnologische Standards zurückgegriffen, deren Anwendbarkeit und Akzeptanz innerhalb der Computer- und Korpuslinguistik zum Zeitpunkt der Antragsstellung noch nicht absehbar war. Durch den rasanten Siegeszug von XML (*extensible markup language*) während der Förderungsphase und die Verwendung von XML in allen Teilprojekten des SFB 441 erschien es sinnvoll, die Abbildung in einen XML-Rahmen einzubetten.

Aus diesem Grund wurde in Steiner (2000) eine XML-Kodierung für das Annotationsschema der TüBa-D spezifiziert. Die Annotation erfolgt über das Tool *Annotate* des SFB 378 (Saarbrücken), für dessen Exportformat eine vollständige XML-Kodierung in Mengel und Lezius (2000) vorliegt. Allerdings inkorporiert die Kodie-

$$
\begin{tikzcd}
\texttt{signature.dtd} \arrow[d, "Instanz"'] & \\
\texttt{signature.xml} \arrow[r, "\texttt{sig2avm.xsl}"] & \texttt{avm.dtd} \arrow[d, "Instanz"'] \\
& \texttt{avm.xml}
\end{tikzcd}
$$

Abbildung 5: Architektur für die XML-Kodierung von AVM-Beschreibungen

rung in Steiner (2000) das Tagset, die möglichen Knoten- und Kantenlabel sowie die strukturellen Eigenschaften des TüBa-Annotationsschemas (z.B. gibt es im Gegensatz zum NEGRA-Schema keine überkreuzenden Kanten, dafür aber uneingebundene Teilbäume bei der Analyse von Parenthesen). Abbildung 4 zeigt die XML-Kodierung der TüBa-D-Analyse des Satzes in (7).

(7) Sarah liebt ein Photo von Peter.

Diese XML-Kodierung diente als Eingabeformat. Das Ausgabeformat sollten XML-kodierte AVM-Beschreibungen sein. In Sailer und Richter (2001) ist ein allgemeines Schema für eine solche Kodierung entwickelt. In Abbildung 5 werden die Komponenten dieses Systems gezeigt. Da alle Signaturen dieselbe Gestalt haben, ist eine DTD für Signaturen vorgegeben (`signature.dtd`). Die konkrete Signatur einer Grammatik ist dann als eine Instanz dieser DTD zu verstehen und als XML-Datei `signature.xml` kodiert. Da die Signatur die Beschreibungssprache genau festlegt, können wir aus der Datei `signature.xml` eine DTD für AVM-Beschreibungen erzeugen. Hierzu verwenden wir XSL-Transformationen, `sig2avm.xsl`. Instanzen der DTD `avm.dtd` sind XML-Kodierungen von AVM-Beschreibungen. Dabei werden die formalen Eigenschaften von RSRL-Grammatiken so ausgenutzt, daß Standard-XML-Parser bereits zur Überprüfung der Wohlgeformtheit einer Beschreibung eingesetzt werden können.

Mit diesen XML-Spezifikationen für Eingabe und Ausgabe konnte die tatsächliche Überführung von Korpusdaten in Strukturskizzen angegangen werden. Im Rahmen der XML-Technologie ist durch die Einführung von XSL (*extensible stylesheet language*) eine leistungsstarke Programmiersprache geschaffen worden, die die Verarbeitung von XML-Dokumenten erlaubt. In ihrer Funktionalität geht XSL weit über die Möglichkeiten von CSS oder DSSSL hinaus, indem in XSL nicht nur relativ einfache, in der Regel ausgabeorientierte Veränderungen definiert werden können, sondern fundamentale Umstrukturierungen an einem XML-Dokument vorgenommen werden können. Die Möglichkeiten von XSL sind so umfassend, daß XSL als Abbildungsformalismus vollständig für die Belange des Projekts ausreichte.

In Abbildung 6 ist das Ergebnis der Konversion des Baums aus Abbildung 4 dargestellt. Aus Platzgründen ist hier eine Baumskizze statt eines AVMs angegeben. Die

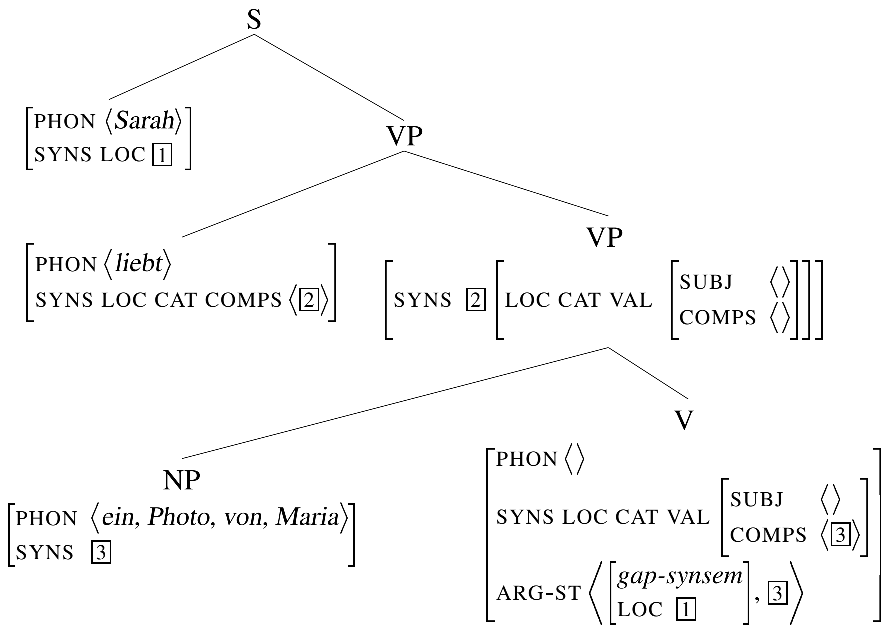

Abbildung 6: Strukturskizze für den Satz in Abbildung 4

eigentliche XML-Ausgabe läßt sich mittels des graphischen Interfaces Grisu (GRaphical vISUalization tool, (Wunsch 2000)) als Klammerstruktur visualisieren. Die Ergebnisbeschreibung ist hinsichtlich einiger Aspekte weniger spezifisch ist als die Eingabestruktur (siehe die Objekts-NP). Andererseits ist sie expliziter in den Bereichen, die für das betrachtete linguistische Phänomen einschlägig sind. So ist die Verbbewegungsanalyse von Meurers (1999) und Tseng (2000) integriert (siehe die Verbspur).

Durch die Klärung des ontologischen Status von theorieneutralen Annotationen und theoriegeleiteten Strukturskizzen konnte unter Verwendung von neuen texttechnologischen Verfahren und Werkzeugen neben dem im Antrag anvisierten Brückenschlag zwischen quantitativen und qualitativen Daten auch ein Brückenschlag zwischen theoretischer Sprachwissenschaft und Texttechnologie erreicht werden.

3 Vergleiche mit Arbeiten außerhalb des Sonderforschungsbereichs und Reaktionen der wissenschaftlichen Öffentlichkeit auf die eigenen Arbeiten

3.1 Automatische inkrementelle Annotation

Für das Englische liegen mittlerweile eine große Anzahl von Untersuchungen vor, die sich besonders mit der Implementierung von maschinellen Lernverfahren zur automatischen Annotation von Chunks befassen (vgl. Daelemans (2000), hier finden sich unter dem Abschnitt *Shared Task Papers* zwölf Arbeiten zu diesem Thema). Daß für das Deutsche vergleichbare Korpora zum Trainieren der Lernverfahren fehlen, ist ein Grund dafür, daß oft regelbasiert annotiert wird. Da die für das Deutsche unten vorgestellten Systeme höchst unterschiedliche Zielsetzungen haben, hat sich noch kein Standard etabliert und die Systeme lassen sich somit nicht direkt vergleichen. Das im Rahmen des SFB erstellte Chunk-Style-Book (Müller 2001) soll die Diskussion über einen solchen Standard anregen.

Das im Rahmen des SFB entwickelte Chunking-System ist unseres Wissens das einzige robuste System, auf dem weitere linguistische Annotation aufbaut. Andere Chunking-Systeme wurden im wesentlichen für Informationserschließungs-Systeme entwickelt und sind daher auch auf diese zurecht geschnitten, was sich unter anderem darin äußern kann, daß sie lediglich Nominal-Chunks und Präpositional-Chunks annotieren. Das Informationserschließungs-System SMES enthält als eine seiner Komponenten ein Chunking-System, das im Rahmen eines flachen Parsing-Systems eingesetzt wird (Neumann, Braun und Piskorski 2000), jedoch auch semantische Elemente enthält. Ebenfalls semantische Elemente enthält das NP/PP-Chunking-System, das Teil des System ResoPPat zur Auflösung von Anbindungsambiguitäten ist (Clematide und Volk 2001). Die von Wauschkuhn (1999) beschriebene Feinanalyse-Komponente seines Systems entspricht einem Chunking-System, arbeitet jedoch nicht robust, da sie auf vorheriger Satzerkennung aufbaut und bei deren Scheitern ebenfalls scheitert.

Zum maschinellen Lernen im Bereich Parsing sind bisher (abgesehen von statistischen Ansätzen) zum größten Teil nur partielle Lösungen diskutiert und implementiert worden. In dem TMR Network "Learning Computational Grammars" wurden verschiedene Ansätze zum Parsing von Nominalphrasen und zum allgemeinen Chunking als gemeinsame Aufgaben implementiert und anhand derselben Trainings- und Testdaten verglichen (siehe z.B. Déjean (2000), Koeling (2000), Osborne (2000), Tjong Kim Sang (2000)). Zu memory-based Verfahren gibt es eine Untersuchung (Argamon, Dagan und Krymolowski 1999), in der Memory-Based Sequence Learning (MBSL) zum NP-Chunking und zur Klassifizierung von Verb-Subjekt- und Verb-Objekt-Relationen verwendet wird. MBSL verwendet Fenster von variabler Länge, um den Eingabestring mit Klammerannotation zu versehen; aus den verschiedenen Klammerungsmöglichkeiten wird dann die konsistenteste mit den höchsten Scores ausgewählt. Es wird jedoch nicht versucht, die unabhängig gelernten Ergebnisse zu einem Parse zu kombinieren. Diese Aufgabenstellung wird bei Buchholz, Veenstra und Daelemans (1999) untersucht. Sie verwenden ein kaskadiertes System, in dem im letzten Schritt die in der Penn Treebank vorhandenen grammatischen Relationen gelernt werden. In beiden Ansätzen können daher keine kompletten Parses für einen Satz ermittelt werden.

3.2 Annotationserschließung

Zu Beginn der Projektphase gab es keine geeigneten Query-Tools um strukturelle Anfragen an ein syntaktisch annotiertes Korpus zu stellen. Durch die zunehmende Verfügbarkeit derartiger Korpora stieg die Notwendigkeit von Tools mit geeigneten Anfragesprachen, um die strukturellen Informationen zu erschließen.

Neuere Anfragesprachen wie z.B. XML-QL (Deutsch, Fernandez, Florescu, Levy und Suciu 1999), die sich auf die XML-Annotation, d.h. auf die Dokumentstrukturen, beziehen, sind nicht geeignet für syntaktisch annotierte Korpora. Diese Sprachen können dann sinnvoll genutzt werden, wenn die Dokumentstrukturen direkt die linguistischen Baumstrukturen repräsentieren. Um diskontinuierliche Konstituenten darstellen zu können, enthalten die meisten syntaktisch annotierten Korpora jedoch von Bäumen abweichende Datenstrukturen (überkreuzende Kanten in der Negra-Baumbank oder unverbundene Knoten in den TüBa-Baumbanken). Die für diese Konstruktionen benötigten XML-Repräsentationen lassen sich nicht mehr auf intuitive Weise direkt durch XML-Anfragesprachen erschließen.

Parallel wurde am IMS Stuttgart (Projekt TIGER) ebenfalls ein Query-Tool für syntaktisch annotierte Korpora entwickelt (siehe Lezius und König (2000), König und Lezius (2000)), welches momentan für die Negra-Baumbank implementiert. Die Anfragesprachen von TIGER und VIQTORYA sind beides Graphbeschreibungssprachen, im Falle von TIGER mit einer operationellen Semantik, im Falle von VIQTORYA mit einer modelltheoretischen Semantik versehen. Auf der eine Seite erlaubt TIGER im Gegensatz zu VIQTORYA ein etwas breiteres Spektrum von Relationen zwischen Knoten und die Definition von Typen und Templates. Dies ändert aber nichts an der Ausdrucksstärke der Anfragesprache, da man die in TIGER zusätzlich möglichen Anfragen auch in VIQTORYA ausdrücken kann, wenn auch auf eine deutlich weniger benutzerfreundliche Weise. Sollte sich herausstellen, daß diese Erweiterungen für linguistisch interessante Anfragen wünschenswert sind, können sie also ohne größeren Aufwand in VIQTORYA integriert werden. Auf der anderen Seite ist TIGER restriktiver als VIQTORYA, da in TIGER nicht nur keine Quantoren sondern auch keine Negation erlaubt ist, während VIQTORYA atomare Negation gestattet. Ob dies eine echte Einschränkung ist, hängt jeweils von den der Korpusannotation zugrundeliegenden Datenstrukturen ab, die ja im Falle von Negra und TüBa verschieden sind. Bei den der TüBa-Annotation zugrundeliegenden Graphen kann man, wie oben schon erwähnt, eine Beziehung `var1 !>> var2` (`var1` dominiert nicht `var2`) für Variablen `var1` $\neq$ `var2`, die verschiedene Knoten bezeichnen, nicht ausdrücken durch `var2 >> var1 | var1..var2 | var2..var1` (`var2` dominiert `var1` oder `var1` steht in linearer Präzedenz zu `var2` oder `var2` steht in linearer Präzedenz zu `var1`). Hier würde ein Verzicht auf atomare Negation also eine echte Einschränkung der Ausdrucksstärke bedeuten. Weiterhin basiert die Architektur von VIQTORYA im Gegensatz zu TIGER auf einer relationalen Datenbank, was eine höhere Performanz des Suchvorgangs erwarten läßt und die Verarbeitung größerer Datenmengen erlaubt.

3.3 Theoriegeleitete Annotationsveränderung

Im Laufe des Antragszeitraums sind etliche fundamentale Arbeiten im Bereich der Grammatikformalisierung entstanden, vor allem hinsichtlich der HPSG. Neben der im Projekt vertretenen Sichtweise zur HPSG-Formalisierung findet sich eine Alternative in verschiedenen linguistischen Aufsätzen, wie beispielsweise in Ginzburg und Sag (1998), und in der Implementierungsplatform LKB (LKB 2000). Dort wird vom informationsbasierten Verständnis von HPSG-Grammatiken, wie es in Pollard und Sag (1987) vertreten wird, ausgegangen.

Im Laufe der Projektphase ist Anzahl HPSG-basierter Baumbanken und Baumbankprojekte gestiegen. So gibt es mitlerweile neben der Baumbank TüBa-E (Kordoni 2000, Flickinger, Copestake und Sag 2000), Projekte für entsprechende Baumbanken für das Polnische (Marciniak, Mykowiecka, Kupść und Przepiórkowski 2000), das Bulgarische (BulTreeBank) und das Niederländische (Alpino, Bouma (2001)). Alle diese Projekte basieren auf Fragmentgrammatiken und sind somit der Gefahr ausgesetzt, entweder nicht robust zu sein, oder linguistisch schwach motivierte Analysen für viele Phänomene zu liefern. Es ist absehbar, daß sich in späteren Phasen diese Projekte die in PB 3 angegangenen Fragen stellen werden.

Während der Förderungsphase wurden für etliche linguistische Formalismen XML-Kodierungen entwickelt und angewendet: für die *Tree Adjoining Grammar* in Bonhomme und Lopez (1999), für relativ allgemeine Baumstrukturen in Mengel und Lezius (2000). Die Arbeitsgruppe von Kiril Simov an der bulgarischen Akademie der Wissenschaften, zu der eine enge Verbindung besteht, beschäftigt sich seit kurzem ebenfalls mit der XML-Kodierung von Datentypen der HPSG. Schließlich sind die Kodierungsverschläge der TEI[5] primär an der Kodierung von Objekten und nicht von Beschreibungen orientiert. Im PB 3 wurde also auch auf der Repräsentationsseite wichtige und für spätere Projekte nutzbare Grundlagenarbeit geleistet.

4 Offene Fragen

4.1 Automatische inkrementelle Annotation

Bezüglich der Ebene des Chunkings stellt sich grundsätzlich die Frage, ob mit den benutzten Methoden und Informationen komplexere Strukturen als bisher annotiert werden können und welche zusätzlichen Wissensquellen dafür nötig wären. Von besonderem Interesse wäre es zu klären, inwieweit topologische Felder und grammatische Funktionen unter Verwendung eines *finite-state*-Ansatzes annotiert werden können und wie zusätzliche Wissensquellen in den *finite-state*-Ansatz integriert werden können.

[5] URL: `http://etext.virginia.edu/TEI.html`

Darüber hinaus muß die Absicherung der Annotationsqualität durch einen weiteren Ausbau des Tag-Verbesserungssystems für die Satzklammerelemente gewährleistet werden. Hierbei muß untersucht werden, inwieweit die bisher positiven Erfahrungen für die Verben auf andere Elemente der Satzklammer wie Konjunktionen und Relativpronomina übertragen werden können.

Beim vollständigen Parsing bleibt die Frage zu beantworten, ob auch hier ein rein inkrementeller Ansatz auf der Baumkonstruktionsebene möglich ist. Werden die grammatischen Funktionen einzeln gelernt, kann die relativ kleine Trainingsmenge, die für syntaktische Annotation zur Verfügung steht, intensiver genutzt werden, da für solche Probleme mehr Instanzen extrahiert werden können. Dadurch ergeben sich auch mehr Arten der Generalisierung. Es bleibt auch zu klären, ob sich die Ergebnisse durch die Verwendung anderer Lernverfahren, wie z.B. durch konnektionistische Methoden, oder durch eine Kombination von Lernverfahren verbessern lassen. Eine weitere Problemstellung ergibt sich auf dem Gebiet der Evaluation. Hier wäre es sinnvoll, Evaluationsmethoden zu entwickeln, die objektivere Bewertungsrichtlinien auch für partielle Parses bereitstellen, oder die eine manuelle Evaluation anhand eines speziell für die Anforderungen ausgewählten Testkorpus möglich machen.

4.2 Annotationserschließung

VIQTORYA ist ein Prototyp und somit existieren vielfältige Möglichkeiten, das Tool weiterzuentwickeln. Insbesondere die Anpassung des Tools an andere Korpora wäre ein wichtiges Ziel. Mit den Projekten A3 und B10 ergaben sich diesbezüglich interessante Diskussionen über Erweiterungen des Query-Tools, d.h. über die Integration von Informationen, die über reine Baumstrukturen hinausgehen, wie z.B. morphologische Information oder sekundäre Kanten. Letztere können verwendet werden für die Kodierung von Koindexierungsrelationen oder temporalen Relationen, wie sie in den Projekten A3 bzw. B10 annotiert werden. Die Anpassung an andere Korpora würde Modifikationen der Initialisierungskomponente bedeuten und in manchen Fällen, wie z.B. bei der oben beschriebenen Erweiterung, würde dies auch zu Modifikationen der Anfragesprache und des Datenbankschemas führen. Hierbei wäre zu untersuchen, inwiefern derartige Erweiterungen Auswirkungen auf das Laufzeitverhalten des Systems haben.

Ein weiterer Schritt wäre, die Einsatzmöglichkeiten des Query-Tools in einem größeren Zusammenhang zu untersuchen, wie z.B. bei der Evaluation von automatischen Annotationswerkzeugen. Insbesondere bei der phänomenorientierten Evaluation, die von einem linguistischen Standpunkt aus attraktiv ist, spielt ein Query-Tool wie VIQTORYA eine wichtige Rolle. So können z.B. alle Sätze, die ein bestimmtes syntaktisches Phänomen aufweisen, mit Hilfe des Query-Tools aus einem annotierten Korpus extrahiert werden, um zu ermitteln, wie gut ein Annotationswerkzeug dieses Phänomen abdecken kann.

4.3 Theoriegeleitete Annotationsveränderung

In diesem Projektbereich wurde in vielerlei Hinsicht theoretisches und praktisches Neuland betreten. Dabei geht es zum einen um die Reflektion des ontologischen Status verschiedener Repräsentationen (Korpus versus Phänokorpus), zum anderen um die Verbindung von texttechnologischen Methoden und formaler Grammatiktheorie. Die zentrale offenen Frage ist, in wieweit sich die hier entwickelten Konzeptionen für vielfältige theoretische und korpuslinguistische Anforderungen eignen.

Phänokorpus: Das Konzept des Phänokorpus als Alternative zu Fragmentgrammatiken und theorieneutraler Annotation muß noch in größerem wissenschaftlichen Kontext propagiert werden. Weitere Anwendungen von Phänokorpora sind im Bereich der Evaluation denkbar, im Hinblick auf phänomenorientierte oder qualitative Evaluation.

Texttechnologie: Die im Projektbereich getroffene Entscheidung zugunsten von texttechnologischen Standards für ein theoriegeleitetes Repräsentationsformat und die Abbildung dorthin steht im Einklang mit der auch in C1 verfolgten Linie. Bedauerlicherweise sind die formalen Eigenschaften dieser Standards noch zu schwach dokumentiert und untersucht. So liefern zwar Berstel und Boasson (2000) eine formale Charakterisierung einfacher XML-Dokumente, für die vollständige Mächtigkeit der XML-Kodierung fehlen entsprechende Arbeiten. Dasselbe trifft auf XSL zu. Hier wäre ein Vergleich mit den Arbeiten zu Baumautomaten und Baumtransformationen notwendig, wie sie im Projekt A4 des SFB 340 durchgeführt wurden und über das Projekt A2 auch auf den SFB 441 Einfluß ausüben.

5 Veröffentlichungen und Manuskripte

5.1 Abgeschlossene Dissertationen

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

Sailer, Manfred (2000): *Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar*, Dissertation, Eberhard-Karls-Universität Tübingen, Version vom 29. Juni 2000. (Teilweise entstanden im SFB 340).

5.2 Veröffentlichungen

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

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

Kallmeyer, Laura (1999*a*): „Synchronous Local TDGs and Scope Ambiguities“, *in* G. Bouma, E. W. Hinrichs, G.-J. Kruijff und R. T. Oehrle, (Hrsg.), *Constraints and Resources in Natural Language Syntax and Semantics*, CSLI, S. 245 – 262.

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

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

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

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

Kallmeyer, Laura (2000*d*): „Scrambling in German and the non-locality of local TDGs“, in *Proceedings of TAG+5 Workshop*, Paris, S. 129 – 134.

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

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

Kübler, Sandra und Andreas Wagner (2000): „Evaluating POS tagging under suboptimal conditions. Or: Does meticulousness pay?“, in *Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications ACIDCA'2000, Tunesien, März 2000.*

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

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

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

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

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

Richter, Frank (1999): „RSRL for HPSG“, *in* V. Kordoni, (Hrsg.), *Tübingen Studies in Head-Driven Phrase Structure Grammar*, Arbeitspapiere des SFB 340, Nr. 132, Band 1, Universität Tübingen, S. 74–115.

Richter, Frank, Manfred Sailer und Gerald Penn (1999): „A Formal Interpretation of Relations and Quantification in HPSG“, *in* G. Bouma, E. Hinrichs, G.-J. M. Kruijff und R. T. Oehrle, (Hrsg.), *Constraints and Resources in Natural Language Syntax and Semantics*, CSLI Publications, S. 281–298.

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

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

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

6 Aktivitäten: Tagungen, Vorträge, Gäste

6.1 Tagungen

Ausrichtung des Internationalen Workshops “Syntactic Annotation of Electronic Corpora”, Schloß Hohentübingen, 13.-15. Juni 2000. Publikation der Tagungsergebnisse als Special Issue in der Zeitschrift “Computational Linguistics” in Vorbereitung.

Zweiwöchiger Kurs “A Comparison of LFG and HPSG” auf der 12th European Summer School in Logic, Language, and Information in Birmingham, England, 6.-18. August 2000.

Einwöchiger Workshop "Linguistic Theory and Grammar Implementation" auf der 12th European Summer School in Logic, Language, and Information in Birmingham, England, 6.-18. August 2000. Publikation der Workshop-Ergebnisse als Special Issue der Zeitschrift "Language and Computation" in Vorbereitung.

Mit-Ausrichtung der Clark-Sommerschule, Sozopol, Bulgarien, 25. August - 8. September 2000.

Mit-Ausrichtung des Workshops "New methods and formalisms for corpus linguistics", Aix-en-Provence, Frankreich, 13.-14. Oktober 2000.

Tutorium "A Mathematical Formalism for HPSG" auf dem CLAIRE Workshop, University of Essex in Colchester, England, 12. Januar 2001.

6.2 Projektübergreifende Aktivitäten

Seminar "Quantitative Grammatikforschung", WS 1998/99, Veranstalter: Erhard W. Hinrichs, Laura Kallmeyer, Marga Reis.

Kurs "Einführung in XML", 13. Oktober 1999, Veranstalter: Lothar Lemnitzer.

Seminar "Statistische Verfahren in der Computerlinguistik", SS 2000, Veranstalter: Erhard W. Hinrichs, Uwe Mönnich.

6.3 Vorträge

Erhard W. Hinrichs: *Linguistic Annotations for Large Corpora of Spoken and Written German* – Einladungsvortrag auf dem Workshop "On the Structure of Non-Narrative Text", University of Texas at Austin, 1999.

Laura Kallmeyer: *Tree Description Grammars and Underspecified Representations* – Vortrag am IRCS, University of Pennsylvania, 29.4.1999

Laura Kallmeyer: *On the Complexity of Queries for Structurally Annotated Linguistic Data* – Vortrag auf der "International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications" (ACIDCA'2000), Tunesien, März 2000.

Sandra Kübler und Andreas Wagner: *Evaluating POS tagging under suboptimal conditions. Or: Does meticulousness pay?* – Vortrag auf der "International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications" (ACIDCA'2000), Tunesien, März 2000.

Laura Kallmeyer: *A query tool for syntactically annotated corpora* – Vortrag auf dem Workshop *Syntactic Annotation of Electronic Corpora* in Tübingen, 13.–15. Juni 2000.

Frank H. Müller und Tylman Ule: *Integrating Multiple Levels of Linguistic Annotation* – Vortrag auf dem Workshop *Syntactic Annotation of Electronic Corpora* in Tübingen, 13.–15. Juni 2000.

Laura Kallmeyer: *A query tool for syntactically annotated corpora* – Vortrag auf der "Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora", Hong Kong, Oktober 2000.

Sandra Kübler und Erhard W. Hinrichs: *Robustes Chunkparsing mit variabler Analysetiefe* – Vortrag auf der KONVENS 2000 / Sprachkommunikation, Ilmenau, Oktober 2000.

Sandra Kübler und Erhard W. Hinrichs: *Robust Chunk Parsing with Tree Construction* – Vortrag auf dem Workshop "New Methods and Formalisms for Corpus Linguistics", Aix-en-Provence, 13.–14. Oktober 2000.

Frank H. Müller: *Using linguistic knowledge to annotate German 'brace constructions' and 'topological fields' in order to reduce ambiguities in overall sentence structure* – Vortrag auf dem Workshop "New Methods and Formalisms for Corpus Linguistics", Aix-en-Provence, 13.–14. Oktober 2000.

Laura Kallmeyer: *Un outil de requête pour des corpus arborés* – Vortrag am TALaNa, Université Paris 7, 30. Oktober 2000.

Frank Richter und Manfred Sailer: *Three kinds of underspecified semantics in HPSG* – Vortrag auf dem "Joint Workshop on Linguistic Data Structures: Theoretical, Methodological, Descriptive Issues" Eberhard-Karls-Universität Tübingen, 19.–20. Dezember 2000.

Erhard W. Hinrichs: *Robust Syntactic Annotation of Corpora and Memory-based Parsing* – Einladungsvortrag auf der Konferenz "Israeli Seminar on Computational Linguistics" (ISCOL'01), Haifa, Israel, 15.2.2001.

Sandra Kübler, Ilona Steiner und Erhard W. Hinrichs: *Automatic Annotation And Querying of German Treebanks* – Vortrag auf der "International Conference on Linguistic Data Structures", Eberhard-Karls-Universität Tübingen, 22.–24. Februar 2001.

Sandra Kübler und Erhard W. Hinrichs: *TüSBL: A Similarity-Based Chunk Parser for Robust Syntactic Processing* – Vortrag auf der "Human Language Technology" Konferenz, San Diego, Cal., März 2001.

Sandra Kübler: *Braucht Nominalphrasenerkennung linguistisches Wissen?* – Vortrag auf der GLDV-Frühjahrstagung 2001, Gießen, 28.–30. März 2001.

Frank H. Müller: *Satzklammer annotieren und Tags korrigieren. Ein mehrstufiges Top-Down-Bottom-Up-System zur flachen, robusten Annotierung von Sätzen im Deutschen* – Vortrag auf der GLDV-Frühjahrstagung 2001, Gießen, 28.–30. März 2001.

Manfred Sailer und Frank Richter: *Eine XML-Kodierung für AVM-Beschreibungen* – Vortrag auf der GLDV-Frühjahrstagung 2001, Gießen, 28.–30. März 2001.

Ilona Steiner: *Warum "Named Entities" für die Chunk-Analyse wichtig sind* – Vortrag auf der GLDV-Frühjahrstagung 2001, Gießen, 28.–30. März 2001.

Sandra Kübler und Erhard W. Hinrichs: *From Chunks to Function-Argument Structure: A Similarity-Based Approach* – Vortrag auf der ACL 2001, Toulouse, Juli 2001.

Frank Richter und Manfred Sailer: *A semantic "Trampe" for HPSG: Easy ways from the bottom to the top and back in Trondheim*, Einladungsvortrag auf der HPSG Conference, Trondheim, August 2001.

6.4 Gastwissenschaftler

Zdravko Peev, LML, Bulgarische Akademie der Wissenschaften, Sofia, 1.11.00 - 28.2.01

Pauline Welby, Ohio State University, Columbus, USA, Juli 2000 (in Kooperation mit B5)

7 Zitierte Literatur

Argamon, Shlomo, Ido Dagan und Yuval Krymolowski (1999): „A Memory-Based Approach to Learning Shallow Natural Language Patterns“, *Journal of Experimental and Theoretical AI* **11**, 369–390.

Berstel, Jean und Luc Boasson (2000): *Formal Properties of XML Grammars and Languages*. Version vom 7. 11. 2000.
URL: *http://www-igm.univ-mlv.fr/~berstel/Articles/MFCSgrammars.ps*

Bonhomme, Patrice und Patrice Lopez (1999): *TagML, version 0.4, Wed Nov 24 1999*.
URL: *http://www.loria.fr/~lopez/TAG_XML/donnees.html*

Bouma, Gosse (2001): „Extracting Dependency Frames from Existing Lexical Resources“, in *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Association for Computational Linguistics, Somerset, NJ.

Brants, Thorsten (2000): „TnT - a statistical part-of-speech tagger“, in *Proceedings of ANLP-2000, Seattle, WA*.

Brants, Thorsten und W. Skut (1998): „Automation of Treebank Annotation“, in *Proceedings of NeMLaP-3/CoNLL98, Sydney, Australia*, S. 49 – 57.

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

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

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

De Kuthy, Kordula (2000): *Discontinuous NPs in German — A Case Study of the Interaction of Syntax, Semantics and Pragmatics*, Dissertation, Universität des Saarlandes, Saarbrücken.

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

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

Feldweg, Helmut (1995): „Stochastische Wortartendisambiguierung des Deutschen“, *in* Helmut Feldweg und Erhard Hinrichs, (Hrsg.), *Lexikon & Text*, Niemeyer, Tübingen, S. 241 – 254.

Flickinger, Dan, Ann Copestake und I. A. Sag (2000): „HPSG Analysis of English“, *in* Wolfgang Wahlster, (Hrsg.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, Springer-Verlag, Berlin, Heidelberg, New York, S. 255–265.

Gallin, Daniel (1975): *Intensional and Higher-Order Modal Logic*, North-Holland, Amsterdam.

Ginzburg, Jonathan und Ivan Sag (1998): *English Interrogative Constructions*. Draft.

Kepser, Stephan (2000): „A Coalgebraic Modelling of Head-Driven Phrase Structure Grammar“, *in* Dirk Heylen, Anton Nijholt und Giuseppe Scollo, (Hrsg.), *Algebraic Methods in Language Processing AMiLP 2000*, S. 81–96.

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

King, Paul J. (1999): „Towards Truth in HPSG“, *in* Valia Kordoni, (Hrsg.), *Tübingen Studies in HPSG*, Arbeitspapiere des SFB 340, Nr. 132, Band 2, Universität Tübingen, S. 301–352.

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

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

Kordoni, Valia (2000): *Stylebook for the English Treebank in* VERBMOBIL, Technischer Bericht 241, Verbmobil.

Lezius, Wolfgang und Esther König (2000): „Towards a search engine for syntactically annotated corpora“, in *Proceedings of the KONVENS Conference*.

LKB (2000): *The (new) LKB System*. System entwickelt von Ann Copestake, John Carroll, Rob Malouf, Stefan Oepen und anderen.
URL: *http://www-csli.stanford.edu/∼aac/doc5-2a.pdf*

Marciniak, Małgorzata, Agnieszka Mykowiecka, Anna Kupść und Adam Przepiórkowski (2000): „An HPSG-Annotated Test Suite for Polish“, in *Proceedings of the Linguistic Resources and Evaluation Conference*.

Mengel, Andreas und Wolfgang Lezius (2000): „An XML-based Encoding Format for Syntactically Annotated Corpora“, in *Proceedings of LREC 2000*, Athen, S. 121–126.

Meurers, W. Detmar (2000): *Spotting strange passives in the news. On the use of electronic corpora for theoretical linguistics*. Vortrag an der Ohio State University, Columbus.

Meurers, Walt Detmar (1999): *Lexical Generalizations in the Syntax of German Non-Finite Constructions*, Dissertation, Seminar für Sprachwissenschaft, Universität Tübingen. Veröffentlicht 2000 als Bericht Nr. 145 in Arbeitspapiere des SFB 340, ISSN 0947-6954/00.

Neumann, Günter, Christian Braun und Jakub Piskorski (2000): „A Divide-and-Conquer Strategy for Shallow Parsing of German Free Texts“, in *Proceedings of ANLP-2000*, Seattle, Washington, S. 239–246.

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

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

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

Pollard, Carl J. (1999): „Strong Generative Capacity in HPSG“, *in* Gert Webelhuth, Jean-Pierre Koenig und Andreas Kathol, (Hrsg.), *Lexical and Constructional Aspects of Linguistic Explanation*, CSLI Publications, S. 281–297.

Richter, Frank (2000): *A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar*, Dissertation, Eberhard-Karls-Universität Tübingen, Version vom 28. April 2000.

Stegmann, Rosmary, Heike Telljohann und Erhard W. Hinrichs (2000): *Stylebook for the German Treebank in* VERBMOBIL, Technischer Bericht 239, Verbmobil.

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

Tseng, Jesse (Hrsg.) (2000): *Aspekte eines HPSG-Fragments des Deutschen*, Arbeitspapiere des SFB 340, Nr. 156, Universität Tübingen.

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

Wunsch, Holger (2000): „Grisu: Grammar Visualization Tool“, unveröffentlichter Programmcode. Universität Tübingen.

Zimmermann, Thomas Ede (1989): „Intensional Logic and Two-Sorted Type Theory“, *The Journal of Symbolic Logic* **54**(1), 65–77.