

Ebenen der Semantikintegration im Textverstehenssystem SYNDIKATE

Inaugural-Dissertation

zur

Erlangung der Doktorwürde
der Philosophischen Fakultäten
der Albert-Ludwigs-Universität
zu Freiburg i. Br.

vorgelegt von

Martin Romacker

aus Karlsruhe

Erstgutachter:	Prof. Dr. Udo Hahn
Zweitgutachter:	Prof. Dr. Rüdiger Klar
Drittgutachter:	Prof. Dr. Jürgen Dittmann
Vorsitzender des Promotionsausschusses des Gemeinsamen Ausschusses der Philosophischen Fakultäten I-IV:	
	Prof. Dr. Peter Auer
Datum der Disputation:	05.02.2001

Danksagung

Bedanken möchte ich mich an erster Stelle bei Herrn Prof. Dr. Udo Hahn (Computerlinguistik, Universität Freiburg), der mich bei dieser Arbeit unterstützt und gefördert hat. In wohl singulärer Weise steht er seinen Doktoranden mit Ratschlägen, kritischen Kommentaren und hilfreichen Hinweisen zur Seite. Mein Dank geht auch an Herrn Prof. Dr. Klar von der Medizinischen Informatik an der Universität Freiburg. Das Projekt, in dessen Kontext diese Arbeit entstand, wurde mit seiner Unterstützung aus der Taufe gehoben. Der Diskurs zwischen der Computerlinguistik und der Medizinischen Informatik hat diese Arbeit sehr vorangebracht. Des weiteren möchte ich mich bei allen Kollegen bedanken, mit denen ich zusammengearbeitet habe. Besonders hervorheben möchte ich Herrn Dr. med. Stefan Schulz, mit dem ich über drei Jahre lang in unserem gemeinsamen Projekt forschen und publizieren durfte.

Ein besonders herzlicher Dank geht an die Redaktionen der Computer-Zeitschriften, die mir freundlicherweise ihre Artikelbestände in maschinenlesbarer Form zur Verfügung gestellt haben. Ohne diese Unterstützung wäre das ansehnliche Korpus zu Texten aus der Informationstechnologie, welches insgesamt ca. 50.000 Artikeln aus über 25 Jahrgängen enthält, niemals entstanden. Insgesamt fünf der 20 angeschriebenen Redaktionen haben uns unterstützt: PC-Welt, PC-Direkt, PC-Professionell, PC-Magazin sowie die Zeitschrift für Computer und Technik ct.

Der Ursprung jeglicher Arbeit sind natürlich stets die eigenen Eltern. Über ihre freundliche Geduld, ihre fortwährende Hilfe und ihr Verständnis habe ich mich immer sehr gefreut.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Die Teildomänen Medizin und Informationstechnologie	5
1.2	Aufbau dieser Arbeit	6
2	Textverstehenssysteme	9
2.1	Informationsextraktionssysteme	13
2.1.1	Das FASTUS-System	13
2.1.2	Das SCISOR-System	14
2.2	Textverstehenssysteme	15
2.2.1	Medical Language Processing	16
2.2.1.1	Der Medical Language Processor des Linguistic String Pro- jects	16
2.2.1.2	Das MEDLEE-System	18
2.2.1.3	Das RECIT-System	20
2.2.1.4	Das MENELAS-System	21
2.2.2	Textverstehen in computerlinguistischen Systemen	22
2.2.2.1	Das KERNEL-System	22
2.2.2.2	Das TACITUS-System	24
2.3	Dedizierte Textverstehenssysteme	25
2.3.1	Zusammenfassung zu Textverstehenssystemen	28
3	Das Textverstehenssystem SYNDIKATE	30
3.1	Anwendungskontext von SYNDIKATE	33
3.2	Architektur von SYNDIKATE: Überblick	34
3.2.1	Grammatik, Parser und Lexikon	36

3.2.2	Wissensrepräsentationssprache, Weltwissen und Semantik	46
3.2.3	Versionierungslerner und Konzeptlerner	54
3.2.4	Skizze des Parsingprozesses	55
3.2.5	Textuelle Phänomene und Anaphernresolution	58
3.2.5.1	Pronominale Anapher	61
3.2.5.2	Nominale Anapher	63
3.2.5.3	Textuelle Ellipse	65
3.2.5.4	Metonymie	68
3.2.5.5	Empirische Validierung zu Textphänomenen	70
4	Semantische Interpretation	73
4.1	Ein schlankes Modell zur semantischen Interpretation	74
4.1.1	Semantisch interpretierbare Subgraphen	76
4.1.1.1	<i>Direct Linkage</i>	85
4.1.1.2	<i>Mediated Linkage</i>	88
4.1.2	Die semantische Ebene	98
4.1.3	Konzeptuelle Interpretation	101
4.2	Ansätze zur semantischen Interpretation	108
4.3	Das Ambiguitätsproblem bei der Wissensextraktion aus Texten	111
4.3.1	Ambiguitäten und das Textverstehenssystem SYNDIKATE	112
4.3.2	Erweiterung der terminologischen Logik um einen uniformen Kontextmechanismus	114
4.3.3	Integration des Kontextmechanismus in SYNDIKATE	118
4.3.4	Behandlung lexikalischer Ambiguitäten	124
4.3.5	Semantische Ambiguitäten bei der Satzanalyse	128
4.3.6	Interpretationskontexte und Disambiguierung	131
4.3.7	Behandlung syntaktischer Ambiguitäten	133
4.3.8	Kontextselektion am Satzende und Interaktion mit der Resolution satzübergreifender Phänomene	136
5	Wissensrepräsentationsformalismen und Domänenwissen	141
5.1	Wissensrepräsentationssprachen und Wissensrepräsentationssysteme	142
5.1.1	Terminologische Zyklen	145
5.1.2	Transitivität von Relationen in terminologischer Logik	147

5.1.3	Temporal Reasoning in Beschreibungslogik	149
5.1.4	Quantifikation und Beschreibungslogiken	152
5.1.5	Wissensrepräsentationssysteme für Beschreibungslogiken	157
5.1.6	Seitenblick auf weitere Wissensrepräsentationssprachen	158
5.2	Ontological Engineering im Kontext der Wissensextraktion aus Texten . .	159
5.2.1	Konzeption und Legitimation des ontologischen Entwurfs in SYNDI- KATE	160
5.2.1.1	Prinzip des sprachunabhängigen ontologischen Entwurfs . .	164
5.2.1.2	Knowledge-Engineering-Prinzipien für Verbkonzepte . . .	166
5.2.1.3	Partonomien in der medizinischen Wissensbasis	169
5.3	Wissensrepositorien	172
5.3.1	Domänenunabhängige Wissensrepositorien	173
5.3.2	Wissensrepositorien in der Medizin	176
6	Evaluierung der semantischen Komponente von SYNDIKATE	182
6.1	Methoden zur Evaluierung der Interpretationsleistungen in Textverstehens- systemen	184
6.2	Extraktion der Datengrundlage aus den Textkorpora	185
6.2.1	Extraktion der Datengrundlage aus IT-Texten	186
6.2.2	Extraktion der Datengrundlage aus medizinischen Texten	189
6.3	Evaluierungsszenario für die semantische Interpretation in SYNDIKATE .	191
6.4	Evaluierung von <i>Direct-Linkage</i> Subgraphen	198
6.5	Evaluierung von <i>Mediated-Linkage</i> Subgraphen	203
6.5.1	Evaluierung von Modal- und Auxiliarkonstruktionen	204
6.5.2	Evaluierung von Präpositionalphrasen	206
6.5.2.1	Klassifikation von Präpositionalphrasen	207
6.5.2.2	Ergebnisse der semantischen Interpretation von Präpositio- nalphrasen	210
6.5.2.3	Qualitative Analyse der empirischen Daten zur Präpositio- nalphrasen	214
6.5.2.4	Qualitativ-quantitative Analyse der empirischen Daten zur Präpositionalphrasen	220
6.6	Abschließende Bewertung zur Evaluation	229

7	Ausblick auf die weitere Entwicklung von SYNDIKATE	232
7.1	Desiderata des Textverstehenssystems SYNDIKATE	232
7.2	Verbesserung und Erweiterung der semantischen Interpretation	233
7.3	Erweiterung und Verfeinerung der empirischen Arbeiten	233
8	Zusammenfassung	236
A	Aufgabendefinition zur Textklassifikation	239
B	Aufgabendefinition zur Klassifikation von Präpositionalphrasen	243
C	Beispieltexte aus dem Trainingsset	248

Tabellenverzeichnis

3.1	Eigenschaften eines Lexikoneintrags	41
3.2	Syntax und Semantik von terminologischen Konzept- und Rollentermen . .	47
3.3	Syntax und Semantik von terminologischen und assertionalen Axiomen . .	49
3.4	Beispiel einer Menge an terminologischen Axiomen	49
3.5	Wohlgeformtheitskriterien für Textphänomene	59
3.6	Centering-Daten für Satz (1) aus dem Textfragment	61
3.7	Centering-Daten für Satz (2) aus dem Textfragment	63
3.8	Centering-Daten für Satz (3) aus dem Textfragment	65
3.9	Centering-Daten für Satz (4) aus dem Textfragment	67
4.1	Syntax und Semantik von kontextabhängigen assertionalen Axiomen	117
6.1	Wahrscheinlichkeitsverteilung bei der Textklassifikation	188
6.2	Statistische Auswertung zur Wortanzahl in IT- und Medizin-Texten	190
6.3	Verstehensleistung bei Genitivattributen	200
6.4	Verstehensleistung bei Modal- und Auxiliarkonstruktionen	205
6.5	Verstehensleistung bei Präpositionalphrasen	212
6.6	Verteilung der Präpositionen in den Trainings- bzw. Testsets	215
6.7	Verteilung der Interpretationsergebnisse auf Präpositionen	228
6.8	Stereotypizität des Sprachgebrauchs	229

Abbildungsverzeichnis

2.1	Information Retrieval versus Information Extraction	10
2.2	Inhaltsrepräsentation eines Texts im MUC-Format	12
3.1	Modelle der Adaption an Subdomänen	31
3.2	Interaktion der Wissensmodule im Textverstehenssystems SYNDIKATE . .	35
3.3	Wortklassenhierarchie mit lexikalischen Knoten als Blättern	40
3.4	Interaktion von Wortaktoren und Phrasenaktoren	42
3.5	Duplizierung von Strukturen bei Ambiguitäten und deren Verwaltung durch Containeraktoren	43
3.6	Dependenzgraph für Satz (1)	45
3.7	Vererbungshierarchie zwischen Wissensbasen	51
3.8	Repräsentation von Satz (1) durch Konzeptgraphen	52
3.9	Interaktion der Wissensquellen bei der Textanalyse (Kontrollfluß)	56
3.10	Satz (2) ohne Resolution der pronominalen Anapher	62
3.11	Sätze (1) und (2) mit resolvierter pronominaler Anapher	62
3.12	Satz (3) ohne Resolution der nominalen Anapher	64
3.13	Sätze (1)-(3) mit resolvierter nominaler Anapher	65
3.14	Satz (4) ohne Resolution der textuellen Ellipse	66
3.15	Sätze (1)-(4) mit resolvierter textueller Ellipse	66
3.16	Satz (5) ohne Resolution der Metonymie	69
3.17	Sätze (1)-(5) mit Resolution der Metonymie	69
3.18	Distribution von Textphänomenen in medizinischen Texten	71
4.1	Dependenzgraph für Satz (6)	75
4.2	Konzeptgraph für Satz (6)	76
4.3	Synthese der Wissens Ebenen bei der statischen Interpretation	80

4.4	Interaktion der Konzeptdefinitionen mit den Restriktionslisten bei der semantischen Interpretation	83
4.5	Semantische Interpretation eines Genitivattributs	87
4.6	Erweiterung der Wortklassenhierarchie um Restriktionsbeschreibungen für die semantische Interpretation an Abhängigkeitsrelationen und intervenierenden Wörtern	90
4.7	Semantische Interpretation einer Präpositionalphrase	91
4.8	Semantische Interpretation mit semantischen Objekten	99
4.9	Beispiel für eine Regel zur konzeptuellen Interpretation	105
4.10	Konzeptuelle Interpretationen in Satz (6)	106
4.11	Kontexte und semantische Interpretation	123
4.12	Beispiel für ambige Wortklassenzugehörigkeit eines Lexikoneintrags	124
4.13	Lexikalische Ambiguität I (Part-of-Speech): Akteure und Initialkontexte . .	125
4.14	Lexikalische Ambiguität II (Polysemie): Akteure und Initialkontexte	127
4.15	Repräsentation semantischer Interpretationsalternativen durch Kontexte .	128
4.16	Repräsentation syntaktischer Ambiguitäten durch Kontexte	134
4.17	Kontextselektion und Kontextkondensierung	138
5.1	Konzeptuelle Repräsentation temporaler Bezüge	151
5.2	Konzeptdefinition für Floppy-Disketten	163
5.3	Reasoning mit Instanzen	164
5.4	Fragment der Ontologie für Verbkonzepte	167
5.5	Reasoning auf partitiven Relationen	170
5.6	Eingeschränkte Generalisierbarkeit partitiver Schlußfolgerungsmuster . . .	171

Kapitel 1

Einleitung

Der Kontext, in dem diese Arbeit entstand, ist das von der Arbeitsgruppe Computerlinguistik an der Universität Freiburg entwickelte und implementierte Textverstehenssystem SYNDIKATE¹ (Hahn & Romacker, 1999b) sowie das von der Deutschen Forschungsgemeinschaft geförderte Projekt *„Automatische Wissensextraktion aus medizinischen Befundtexten“*. Damit erweitert sich die Themenstellung von rein computerlinguistischen Fragestellungen in Richtung der Probleme und Methoden, die in der Medizinischen Informatik relevant sind.

Im Vergleich zu anderen Textverstehenssystemen weist SYNDIKATE einerseits eine Reihe von Standardkomponenten auf, andererseits verfügt SYNDIKATE aber auch über eine Anzahl von besonderen Schlußfolgerungs- und Wissenskomponenten, die in ihrer Kombination noch in keinem vergleichbaren System zur Anwendung kommen. Obwohl sich Kapitel 3 ausführlich mit der Systemarchitektur beschäftigt, seien an dieser Stelle bereits die Analyse referentieller Bezüge oberhalb der Satzebene, die unmittelbare Integration figurativer Sprache in den Textanalyseprozeß sowie die kontinuierliche Erweiterung des Weltwissens durch Konzeptlernen angeführt. Der Kern dieser Arbeit befaßt sich indessen mit der Transformation syntaktischer Beschreibungen, in SYNDIKATE in Form von Dependenzgraphen, in eine inhaltliche Repräsentation auf der Grundlage einer formalen Wissensrepräsentationssprache. Diese Abbildungsoperation von syntaktischen Strukturen auf inhaltliche Repräsentationen bezeichnet man im allgemeinen als semantische Interpretation (vgl. Allen (1997)). Ausgehend von der Tatsache, daß dependentielle Syntaxbeziehungen

¹SYNDIKATE ist ein Akronym für SYNthesis of DiStributed Knowledge Acquired from Texts.

eine strukturelle Ähnlichkeit zu Weltwissens-Prädikaten besitzen (Hajicova, 1987), erfolgt die semantische Interpretation unmittelbar auf den berechneten Weltwissensstrukturen, d.h. ohne intermediäre Repräsentationsschicht, indem eine limitierte Anzahl von formal beschriebenen Graph-Konfigurationen innerhalb eines Abhängigkeitsgraphen adressiert wird.

Grundlage der sprachwissenschaftlichen Betrachtungen in dieser Arbeit sind authentische Texte, so wie sie in einem randomisierten Verfahren aus den vorhandenen umfangreichen Textkorpora extrahiert wurden (vgl. Kapitel 6). Eine redaktionelle Überarbeitung von Eingabetexten findet demzufolge nicht statt. Damit rücken in das Zentrum der formalen Beschreibungen nicht etwa syntaktische oder semantische Idiosynkrasien, sondern die in den Texten enthaltenen hochfrequenten Problemklassen. Ebenso wichtig wie der Gesichtspunkt der Authentizität des Textmaterials ist die Anwendungsorientierung, die natürlich mit dem Authentizitätsaspekt verknüpft ist. D.h., Ziel der Wissensakquisition aus Texten ist nicht die Behandlung semantischer Nuancen und Konnotationen, sondern die Extraktion einer maximalen Anzahl von in den Texten enthaltenen Propositionen in Form von Faktenaussagen, aber auch Bewertungen oder vergleichenden Aussagen. Daß mit der höheren Gewichtung der breiten inhaltlichen Abdeckung gegenüber einer reichhaltigen semantischen Repräsentation natürlich auch ein Teil des Textinhaltes unerschlossen bleibt, soll hier nicht verhehlt werden.

Der explizite Anwendungsbezug von SYNDIKATE betont die Frage nach den tatsächlich automatisierbaren Verfahren bei der Wissensextraktion und kann nicht bei einer rein formalen Beschreibung stehen bleiben. Vielmehr müssen die konkreten Anforderungen an das Textverstehen auch in der Implementation umsetzbar sein. Ein Aspekt, der beispielsweise bei der Wissensrepräsentation eine entscheidende Rolle spielt, besteht darin, daß die formale Eigenschaft der Ausdrucksmächtigkeit einer logischen Beschreibungssprache gegen die Berechnungskomplexität des im System zur Anwendung kommenden Wissenrepräsentationssystems austariert werden muß (vgl. Kapitel 5.1). Auch hier liefert der empirische Befund, wie häufig Klassen von Phänomenen in den Textkorpora auftreten, die notwendigen Indikatoren für die Präferenzen, um die adäquaten formal-sprachlichen Beschreibungsmittel auszuwählen. Als Beispiel kann die Wichtigkeit partonymischer Beziehungen zur Beschreibung anatomischer Konzepte in der Medizin dienen (vgl. Kapitel 5.2.1.3).

Damit verschiebt sich der Interessensfokus von einer ausschließlich am Logik-Paradigma orientierten, rein formalen Beschreibung des Inhalts eines Eingabetext hin zu den infe-

rentiellen Mechanismen, *wie* aus einem Eingabetext inhaltliche Beschreibungen in Form logischer Axiome zu gewinnen sind.² Darüber hinaus definiert der Anwendungsbezug die Informativität von Wissen, das in einem Text verborgen liegt und welches es mithilfe computergestützter Textanalyse zu extrahieren gilt. D.h., daß beispielsweise in der Medizin der Extraktion von Staging- und Gradingaussagen³ über Erkrankungen ein höherer Stellenwert zukommt als etwa der Skopusbestimmung bei definiten Nominalphrasen, da das Anwendungsinteresse in einer automatischen Abbildung von Freitexten auf die diskreten Werte der Staging- und Gradingsskalen liegt.⁴ Damit wird festgelegt, *welche* Aussagen überhaupt während der Textanalyse aus den Eingabetexten gewonnen werden müssen. Daß der von uns verwendete Ansatz zur Faktenextraktion trotz seiner Anwendungsorientierung eine formale Fundierung in der Logik besitzt (hier in einer terminologischen Logik, vgl. Kapitel 3.2.2), gehört zu seinen Charakteristika.

Wenn die stark formal-logisch ausgerichtete Diskussion semantischer Interpretationen den einen Pol markiert, gegenüber dem der Ansatz von SYNDIKATE abzugrenzen ist, dann ist der andere Pol durch Systeme zur Informationsextraktion gegeben. Bei der *Bewertung* dieser Systemen stehen nicht so sehr die zugrundeliegenden Verfahren oder die Formalisierung im Vordergrund, sondern die beim Textanalysieren erzielte Verstehensleistung, welche im Systemvergleich mithilfe eines objektiven Verfahrens ermittelt wird (vgl. Kapitel 2). D.h., neben ihrem Stellenwert für die formal-linguistische Theoriebildung reflektieren sie insbesondere den Stand der Kunst bei der automatischen Wissensextraktion auf großen Textmengen. Die Verstehensleistung kann jedoch mit der von SYNDIKATE nicht verglichen werden, da sich Systeme zur Informationsextraktion von vorneherein auf eine begrenzte Anzahl prädefinierter Templates beschränken, die einen erheblichen Teil der Textinhalte gar nicht zu repräsentieren in der Lage sind. Darüber hinaus fällt die ei-

²Als beispielhafter Repräsentant für eine stärker logisch-formale Betrachtung der Textanalyse sei die Arbeit von Paredes-Frigolett & Strube (1996) angeführt. Beim inkrementellen Aufbau der logischen Axiome, die der inhaltlichen Repräsentation von Eingabesätzen dienen, blenden sie die Frage, wie die logischen Axiome aus der Eingabe *berechnet* wurden, gänzlich aus. Im Vordergrund steht das Interesse am beschreibenden formalen System.

³Staging und Grading spielen eine wichtige Rolle bei der Klassifikation von Erkrankungen, indem sie deren Fortschreiten und Schwere beschreiben.

⁴Dies läßt sich dadurch belegen, daß in der nahezu vollständigen Übersicht zu Textverstehenssystemen in der Medizinischen Informatik von Spyns (1996) keine Arbeit enthalten ist, die sich mit Skopusproblemen beschäftigt. Jedoch extrahieren die meisten der Systeme Graduierungsaussagen.

gentliche Repräsentation in Form einer frameartigen Kombination von Entitäten im Text relativ flach aus. Im Spannungsfeld zwischen den skizzierten Polen der *Language Theory Approaches* einerseits und der *Language Engineering Approaches* andererseits läßt sich SYNDIKATE verorten. In Analogie zu den Informationsextraktionssystemen verarbeitet SYNDIKATE authentische Eingabetexte, erzielt im Vergleich zu diesen ein im Prinzip tieferes, weil a priori weniger restringiertes Textverstehen. Freilich weist das aus den Texten extrahierte Wissen in seinen Beschreibungen nicht die semantische Tiefe der formal-logisch orientierten Ansätze auf, denn beispielsweise Quantifikation wird beim augenblicklichen Stand der Implementation von SYNDIKATE nicht berücksichtigt. Diese formal-logischen Ansätze wiederum kümmern sich indessen weniger um Probleme der realistischen Textanalyse und führen keine empirischen Studien durch, die ihre tatsächliche Performanz an authentischen Texten messen.

Die angestrebte tiefe inhaltliche Erschließung von Texten durch SYNDIKATE erfordert die Integration heterogener Wissensquellen während des Analyseprozesses. Da die Berechnung syntaktischer Strukturen gemessen an diesem Ziel nur einen intermediären Schritt beim Aufbau der eigentlichen Inhaltsrepräsentation bedeutet, erfolgen die verstehensrelevanten Berechnungen in der semantischen Ebene. Die semantische Ebene muß außerdem die Resolution der beim Textverstehen auftretenden referentiellen Bezüge in der Textwissensbasis vornehmen.

Im Gegensatz zu der ebenfalls in der Arbeitsgruppe Computerlinguistik in Freiburg entstandenen Dissertation zum dependenzgrammatischen Ansatz von Norbert Bröker (Bröker, 1999) konzentriert sich daher diese Arbeit bei der Integration heterogener Wissensquellen auf die Darstellung der semantischen Interpretationen, die eng mit den syntaktischen Beschreibungen verbunden ist, wie auch der theoretischen und praktischen Fragestellungen bei der Wissensrepräsentation. Außerdem bietet sie einen umfassenden Überblick über das Gesamtsystem und veranschaulicht die Interaktion der Systemkomponenten bei der Textanalyse.

1.1 Die Teildomänen Medizin und Informationstechnologie

Die in dieser Arbeit verwendeten Textkorpora entstammen zwei sehr unterschiedlichen Diskurswelten, nämlich der Medizin einerseits und der Informationstechnologie andererseits. Auch ohne in diesen Weltausschnitten über die notwendige begriffliche Expertise zu verfügen, ist unmittelbar einsichtig, daß der sich überlappende Bereich an Weltwissen nur einen Bruchteil des jeweiligen Spezialwissens ausmacht. Anatomische Konzepte sind in der Informationstechnologie sicherlich genauso irrelevant wie die Hierarchie der Prozessorfamilien in der Medizin. Gleichwohl bringt die enorme konzeptuelle Divergenz der Teildomänen einen wichtigen methodischen Gewinn mit sich. Sie verhindert, daß die im Textverstehenssystem verwendeten Wissensmodule nur eine hochgradig adaptierte Partiallösung für einen relativ begrenzten Weltausschnitt darstellen. Der anvisierte hohe Grad an Wissensabstraktion gewährleistet die Portabilität der Textverstehenssystems SYNDIKATE auf andere Domänen und untermauert damit die zur Anwendung kommenden Entwurfsprinzipien.

Ebenfalls begleitet haben diese Arbeit die unschätzbaren Reize, aber auch Gefahren der Interdisziplinarität zwischen Linguistischer Informatik auf der einen Seite und Medizinischer Informatik auf der anderen Seite. Innerhalb der Medizinischen Informatik hat sich die maschinelle Verarbeitung natürlichsprachlicher Texte als eigenständiger und anerkannter Forschungs- und Anwendungsbereich etabliert (eine Übersicht zu derartigen Systemen findet sich in Spyns (1996)). Auf den ersten Blick mag überraschen, daß im Bereich der Verarbeitung natürlicher Sprache (NLP) das über den längsten Zeitraum vorangetriebene Projekt nicht etwa in der eigentlichen Computerlinguistik angesiedelt ist, sondern in der Medizinischen Informatik. Seit den frühen siebziger Jahren wurde kontinuierlich innerhalb des *Linguistic String Projects* an der automatischen Analyse medizinischer Freitexte gearbeitet (siehe Sager et al. (1987)). Jedoch waren und sind die Projektarbeiten zu NLP in der Medizinischen Informatik i.a. im Kontext konkreter Applikationen oder zumindest eines konkreten Anwendungs- wie auch mitunter kommerziellen Verwertungsinteresses angesiedelt (Lyman et al., 1991). Dementsprechend gibt es in der Medizinischen Informatik gar kein bzw. nur wenig Interesse an linguistischer Theoriebildung, vielmehr werden Systeme zur automatischen Verarbeitung natürlicher Sprache an ihrer Performanz in Applikationsszenarien gemessen. Hieraus ergibt sich unmittelbar die Notwendigkeit, die NLP-Systeme zu evaluieren. Aus diesem Grund fin-

den sich eine Reihe methodischer Reflektionen, wie NLP-Systeme umfassend zu evaluieren sind, im Kontext der Medizinischen Informatik (vgl. Friedman & Hripcsak (1998), Zweigenbaum et al. (1997)). Die Forderung, Ansätze zur Analyse natürlicher Sprache zu evaluieren, markiert allerdings einen Trend, der auch zunehmend die Computerlinguistik bestimmt.

Ein wenig verwunderlich erscheint die Tatsache, daß zwischen der an der computergestützten Analyse natürlicher Sprache interessierten Medizinischen Informatik und der Computerlinguistik nahezu kein Theorie- und Methodentransfer stattfindet, sondern beide wissenschaftliche Gemeinschaften nahezu isoliert nebeneinander existieren. Die Publikations- und Konferenzplattformen sind mehr oder weniger disjunkt. Unter diesen Voraussetzungen gewinnt der Aspekt der Interdisziplinarität an zusätzlichem Gewicht, indem man sich einerseits den Anwendungsbedürfnissen der Medizinische Informatik stellt und andererseits eine saubere linguistische Theoriebildung, wie sie die Computerlinguistik fordert, gewährleistet. In diesem Sinn können die im Rahmen des interdisziplinären DFG-Projekts erschienenen Publikationen einen vermittelnden Beitrag zwischen den beiden Wissenschaften leisten.

1.2 Aufbau dieser Arbeit

Das Kapitel 2 versucht zunächst, verschiedene methodische Zugänge zum Textverstehen aufzuzeigen. Dabei wird zunächst das *Information Retrieval* von *Information Extraction* unterschieden. Im Anschluß daran kommen einige Informationsextraktionssysteme bzw. Textverstehenssysteme zur Darstellung. Entsprechend der interdisziplinären Perspektive werden hierbei sowohl Systeme aus der medizinischen Informatik als auch aus dem computerlinguistischen Kontext berücksichtigt.

Im Anschluß daran gibt das Kapitel 3 einen umfassenden Überblick über SYNDIKATE, der sich sowohl mit dem eigentlichen Anwendungskontext von SYNDIKATE auseinandersetzt als auch die einzelnen Wissenmodule umreißt und deren Interaktion beim Textverstehen verdeutlicht. Den beim Textverstehen aufzulösenden Referenzphänomenen kommt in der Darstellung ein besonderes Gewicht zu, da die Tiefe der Verstehensleistung wie auch die Validität, die Konsistenz und die Kohärenz der zu generierenden Textwissenbasen von der Resolution anaphorischer Phänomene abhängt.

Im Kapitel 4 wird der in SYNDIKATE zur Anwendung kommende schlanke Ansatz

zur semantischen Interpretation aufgezeigt. Da die Ausgabe des Textverstehenssystems in einer inhaltlichen Repräsentation eines Eingabetexts besteht, stellt die Abbildung syntaktischer Bezüge auf eine Inhaltsrepräsentation – wie sie die semantische Ebene leistet – die eigentliche Integrationsleistung bei der Koppelung heterogener Wissensquellen dar. Einen weiteren thematischen Schwerpunkt im Kapitel zur semantischen Interpretation bildet die Behandlung unterschiedlicher Ambiguitätsebenen mithilfe eines Kontextmechanismus. Seine besondere Anziehungskraft gewinnt der hierzu eingeführte Ansatz einerseits durch eine stringente Trennung der Wissensebenen bei der Repräsentation von Mehrdeutigkeiten und andererseits durch seine formale Fundierung in einem logikbasierten Formalismus.

Der in SYNDIKATE verfolgte Ansatz zum Textverstehen ist stark wissensbasiert. Den Prozessen bei der semantischen Interpretation liegt insbesondere detailliertes Allgemeinwissen, aber auch domänenspezifisches Wissen zugrunde. Für die maschinelle Verwendbarkeit muß dieses Wissen in einer formal fundierten Wissensrepräsentationssprache vorliegen. Jedoch gilt es zu klären, welche (minimalen) Anforderungen das Textverstehen an die Ausdrucksmächtigkeit einer Wissensrepräsentationssprache stellt. Darüber hinaus erfordern Textverstehenssysteme, die für ein weiteres Einsatzfeld konzipiert sind, ein umfangreiches “Weltwissen” bestehend aus mehreren zehntausend Konzepten mit reichhaltigen Eigenschaftsbeschreibungen. Die Frage lautet daher: Wo bieten sich Ansatzpunkte, bestehende Begriffssysteme, Vokabularien oder Wissensrepositorien für die Verwendung in einem Textverstehenssystem auszunutzen? Das Kapitel 5 nimmt sich dieser Fragestellungen an und diskutiert dabei den verwendeten Wissensrepräsentationsformalismus.

Mit dem Kapitel 6 geht diese Arbeit über die formale Betrachtung des entwickelten schlanken Ansatzes zur semantischen Interpretation hinaus. Dieser in Kapitel 4 eingeführte Ansatz wird für drei Klassen von Sprachphänomenen einer Evaluation in beiden Anwendungsdomänen unterzogen. Das Ausgangsmaterial der Evaluation besteht aus einer Untermenge an Texten, die in einem randomisierten Verfahren aus unseren Textkorpora extrahiert wurde. Da sich die beiden Varianten von SYNDIKATE, nämlich die für die Informationstechnologie und die für die Medizin, auf jeweils eine Textsorte konzentrieren, Produktberichte bzw. histologische Pathologieberichte, besteht das eigentliche Testset nur aus Texten, die in einem empirischen Klassifikationstest diesen beiden Textgruppen zugeschlagen wurden. An dieser Stelle sei bereits angemerkt, daß die in diesem Kapitel vorgestellten Arbeiten nur einen Schritt in die richtige Richtung markieren. Insbesondere das Fehlen eines verbindlichen Goldstandards in Form semantisch annotierter Textkorpora

macht die eigentliche Bewertung der Ergebnisse, die die inhaltliche Repräsentation von Eingabetexten darstellen, zu einem noch zu lösenden methodischen Problem.

Abgeschlossen wird diese Arbeit im Kapitel 7 mit einer kleinen Aufzählung von Perspektiven, in welche Richtungen das Textverstehenssystem SYNDIKATE weiterzuentwickeln ist. Als Endperspektive soll SYNDIKATE in realistischen Anwendungsszenarien zum Einsatz kommen. Im Kontext eines Klinikinformationssystems könnten beispielsweise die routinemäßig im Klinikalltag anfallenden Freitexte inhaltlich durch SYNDIKATE erschlossen werden, um dann etwa für faktengestützte Information-Retrieval-Dienste oder automatische Befundverschlüsselung zur Verfügung zu stehen. Bevor dieses ambitionierte Ziel erreicht werden kann, sind indessen noch Erweiterungen erforderlich. Diese Erweiterungen betreffen zwei Ebenen. Erstens erfordert der Einsatz in einem breiteren Rahmen die Skalierung des Systems im Sinn einer rein quantitativen Füllung bestehender Spezifikationen in allen Wissensmodulen. Dazu gehört der Ausbau des Lexikons, die Erweiterung des Domänenwissens wie auch des domänenunabhängigen Wissens oder die Vervollständigung der grammatikalischen Beschreibungen. Zweitens bedürfen jedoch noch eine Reihe von Problemen aufgrund ihres prominenten Auftretens beim Sprachverstehen einer grundsätzlichen Lösung. In diesen Bereich gehören u.a. die Koordination sowohl auf der syntaktischen wie auch auf der semantischen Ebene oder die Differenzierung nach Individuen, Mengen von Individuen, Mengen von Konzepten sowie Konzepten und Konzeptklassen bei der Projektion von Diskursentitäten auf Weltwissensstrukturen. Allen hier anklingenden Defiziten zum Trotz erzielt SYNDIKATE gegenüber klassischen Informationsextraktionssystemen ein tieferes Textverstehen. Denn grundsätzlich wird angestrebt, jede in den Texten anzutreffende Phrase zu interpretieren. Anschließend werden die auf Phrasenebene berechneten Teilergebnisse kompositional zu größeren Interpretationseinheiten kombiniert. Die nächsten beiden Kapitel verdeutlichen diesen Zusammenhang.

Kapitel 2

Textverstehenssysteme

Die Notwendigkeit einer inhaltlichen Erschließung von Texten nimmt in dem Maße zu, wie sich der Zugriff unterschiedlicher Nutzergruppen auf größere Bestände an in elektronischer Form verfügbaren Texten ausweitet. Durch diese Entwicklung verschärft sich das Informationsbeschaffung- und Informationsmanagementproblem zusehends. Die global gesehen mit Sicherheit größte Ressource stellt mittlerweile das Internet mit seinen vielen Millionen Seiten dar. In der Medizin kommt darüber hinaus noch eine immense Anzahl an nicht-öffentlichen Dokumenten hinzu, die im Klinikalltag anfallen und in den Informationssystemen der Klinikrechenzentren abgespeichert sind (etliche GigaByte Daten pro Jahr in mittleren Hospitälern, siehe Hersh et al. (1996)). Gemeinsam haben die meisten dieser Texte, daß die Informationen in ihnen unstrukturiert in völlig freier Form vorliegen. Zur inhaltlichen Erschließung dieser Texte existieren unterschiedliche Methoden.

Das klassische Information Retrieval (IR) erfaßt alle Dokumente einer Textkollektion durch eine Reihe unterschiedlicher Indexierungsverfahren (vgl. hierzu z.B. Salton et al. (1994) sowie für die Medizin Hersh (1996)). Damit werden diese Texte jedoch nicht inhaltlich erschlossen, sondern nur auf Indexterme abgebildet. Linguistisches Wissen kommt hierbei nur in einem bescheidenen Grad zum Einsatz, zum Beispiel durch die Eliminierung von Stoppwörtern oder durch einfache Stemming-Algorithmen, die Flexionssuffixe abschlagen. Auf diese Weise ermöglichen IR-Systeme das Auffinden relevanter Dokumente in einer Textkollektion durch eine vom Benutzer eingegebene, aus Indextermen bestehende Suchanfrage, die über Boole'sche Operatoren und Metadeskriptoren weiter verfeinert werden kann (Abb. 2.1 oben). Im Anschluß an das Retrieval muß jedoch der Nutzer die gefundenen Texte sichten und sich deren Inhalt intellektuell erschließen.

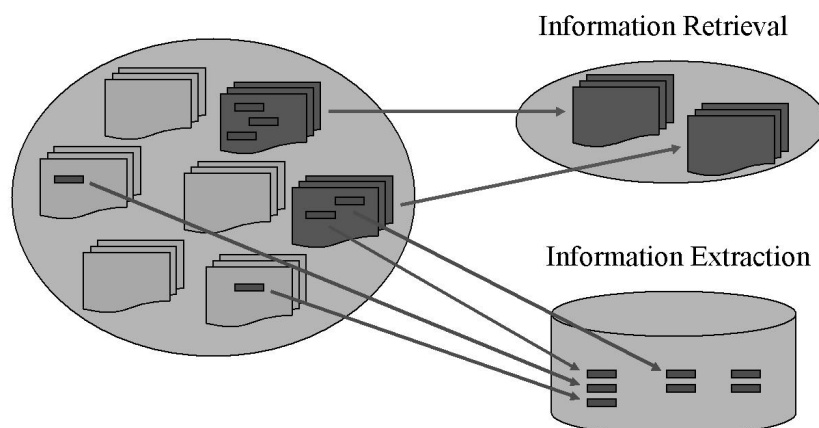


Abbildung 2.1: Information Retrieval versus Information Extraction

Informationsextraktionssysteme (IE) hingegen extrahieren automatisch die Inhalte eines Eingabetextes, indem sie in einem Text enthaltene Aussagen auf Slots an vordefinierten Informationstemplates abbilden und mit diesen inhaltlichen Repräsentationen den Zugriff auf die in den Texten enthaltenen Aussagen ermöglichen (Abb. 2.1 unten). Vergleichende wissenschaftliche Untersuchungen zu den Verstehensleistungen von Informationsextraktionssystemen werden in unregelmäßiger Folge im Zusammenhang der MESSAGE UNDERSTANDING CONFERENCES (MUC) vorgenommen. Die in diesen Vergleichsstudien konkurrierenden Systeme müssen eine kleinere Anzahl an unbekannten, jedoch auf begrenzte Domänen (z.B. Terrorangriffe in Südamerika) und auf einfachere Textsorten (z.B. Zeitungsartikel) beschränkte Texte analysieren, ohne daß im Anschluß an eine Trainingsphase irgendwelche Erweiterungen oder Adaptionen am System vorgenommen werden dürfen. Die Evaluation der von den IE-Systemen produzierten Ergebnisse erfolgt mit Hilfe eines durch manuelle Annotation der Eingabetexte geschaffenen Goldstandards, der zur Ermittlung der Recall- und Precisionwerte dient (vgl. Grishman & Sundheim (1996a)).¹ Dabei hat die Komplexität der Extraktionsaufgaben in den letzten Jahren kontinuierlich zugenommen (Grishman & Sundheim, 1996b). War zu Beginn in MUC-3 die Anzahl der Slots,

¹Der Recallwert bezeichnet die Anzahl der relevanten Dokumente in der Ergebnismenge zu einer Anfrage im Verhältnis zur Gesamtzahl aller relevanten Dokumente in einem Textkorpus. Der Precisionwert gibt an, in welchem Verhältnis die gefundenen relevanten Dokumente zur Gesamtzahl der insgesamt gefundenen Texte steht. Die Evaluationsthematik kommt in Kapitel 6 ausführlich zur Darstellung.

die in den Templates für Ereignisse ausgefüllt werden mußten, noch auf 18 begrenzt (Sundheim, 1991), so stieg sie auf 24 Slots für die Textverstehensaufgabe in MUC-4 (Sundheim, 1992). Jedoch brachte erst MUC-5 (Sundheim, 1994) eine tatsächliche Erweiterung der Aufgaben zur Informationsextraktion, indem die strenge Templatestruktur, die pro Slot nur ein atomares Attribut zuließ, zugunsten einer etwas komplexeren Objektsprache aufgegeben wurde. Damit wurden sowohl multiple Instantiierungen für Slot-Belegungen als auch eingebettete Objektbeschreibungen ermöglicht (Sundheim, 1995). Schließlich erweiterte man das Spektrum an Verstehensleistungen auf die erfolgreiche Disambiguierung von Wortbedeutungen, die Berechnung von Prädikat-Argument Strukturen und sogar auf die Erkennung von Koreferenzbeziehungen zwischen Nominalphrasen, die über eine Identitätsrelation repräsentiert werden. Interessanterweise favorisiert Sundheim (1995) im Rahmen der für MUC-6 formulierten Systemziele eine Reihe von Eigenschaften, die wichtige Designprinzipien von SYNDIKATE ausmachen (Grishman & Sundheim, 1996a): domänenunabhängiger, modularer Aufbau von IE-Systemen, Portabilität und "tieferes" Verstehen (vgl. hierzu Kap. 3). In MUC-7 (1998) schließlich kam als zusätzliche Extraktionsleistung die Relationierung zwischen Templates hinzu. Damit erweiterte sich das Spektrum auf folgende Aufgaben (Chinchor, 1998):

- Erkennung benannter Entitäten (*Named Entities*), d.h. Erkennung von Personen, Organisationen, Orten, Daten, Zeitangaben, Prozentangaben und Geldbeträgen;
- Erkennung von Attributen von Elementen eines prädefinierten Musters (*Template Elements*) für eine Reihe von Slots wie Name, Typ, Deskriptoren;
- Erkennung von Koreferenzbeziehungen bei der Verwendung von Proformen und Umschreibungen, wodurch benannte Entitäten und *Template Elements* miteinander verbunden werden;
- Relationierungen zwischen *Template Elements*, sogenannte *Template Relations*, die die aus den Nachrichtentexten extrahierten Fakten repräsentieren;
- *Events*, die in Form eines Szenario-Templates dargestellt werden, vereinigen die Template-Elemente in sich.

Die erzielte inhaltliche Repräsentation orientiert sich an einer frameartigen Darstellung (vgl. Minsky (1975)). Zur Illustration, welche Repräsentationsstruktur in MUC-7 zur

```
<employee_of-2>
Person: <Entity-3>
Organization: <Entity-4>

<Entity-3>
Entity_Name: "Dennis Gillespie"
Entity_Type: Person
Entity_Description: "Wing Commander"

<Entity-4>
Entity_Name: "NAVY"
Entity-Type: Organization
Entity-Category: Governmental_Organization
```

Abbildung 2.2: Inhaltsrepräsentation eines Texts im MUC-Format

Darstellung eines Faktums Verwendung finden, sei ein Beispiel angeführt (siehe Abb. 2.2). In diesem Beispiel verbinden sich zwei *Named Entities* (Entity-3 und Entity-4), die eine Reihe von Attribute wie Entity_Type oder Entity_Description aufweisen, über eine *Template Relation* (employee-of) miteinander (in Anlehnung an Chinchor (1998)).

Informationsextraktionssysteme operieren demnach auf einem klaren Informationsmodell, d.h., die aus einem Text zu gewinnenden Entitäten, Fakten und Relationierungen sind im voraus eindeutig festgelegt. Daraus folgt, daß trotz mittlerweile stark erweiterter Anforderungen an die Verstehensleistung eine dennoch eingeschränkte Analysefähigkeit verlangt wird, die überdies an begrenzte Domänen adaptiert ist. Ein Problem der methodischen Beschränkungen bei der Textwissensextraktion, wie MUC sie vorgibt, besteht darin, daß Änderungen an der Template-Struktur eine komplette Reanalyse der gesamten Textkollektion mit sich bringen.

Im Gegensatz zu Textverstehenssystemen diese Typs propagieren wir mit SYNDIKATE eine zunächst offene, im Prinzip unrestringierte Form des Textverstehens, die sich durch die Tiefe der Wissensakquisition aus Texten auszeichnet, ohne dabei auf prädefinierte Informationstemplates beschränkt zu sein (also im Sinne von Palmer et al. (1993) ein *broad range, multiple purpose* System). Damit ist gemeint, daß aus einem Eingabetext zunächst

eine maximale Anzahl einfacher Fakten gewonnen wird (Beispiel: “der Durchmesser des Biopsiepartikels beträgt *4 mm*”). Grundsätzlich steht also auf Phrasenebene die Erschließung sämtlicher Bestandteile im Text im Vordergrund, die inhaltsrelevant sind. Jedoch sollen darüber hinaus komplexe Propositionen (Beispiel: “In *allen* Gewebeproben aus dem subkutanen Fettgewebe fanden sich Lymphknotenmetastasen eines undifferenzierten Karzinoms”) sowie – in der Medizin besonders wichtige – evaluierende Aussagen (Beispiel: “Eine *mittelgradige* chronische Duodenitis”) extrahiert werden. Ein weiteres Merkmal von SYNDIKATE besteht in der Integration von Textphänomenen in den Analyseprozeß (vgl. Kap. 3.2.5). Ein Text ist eben nicht nur die bezugslose Aneinanderreihung von Sätzen. Vielmehr ergibt sich eine valide und vollständige inhaltliche Repräsentation des Textinhaltes durch die Resolution von Bezügen zwischen Sätzen (lokale Kohärenz) und Textabschnitten (globale Kohärenz).

Nachdem im nächsten Kapitel 2.1 zunächst zwei Beispiele für Informationsextraktionssysteme gegeben werden, führt das daran anschließende Kapitel 2.2 in vom Ansatz her weiter gefaßte Textverstehenssysteme ein. Die hierzu gehörenden Unterkapitel 2.2.1 und 2.2.2 dienen dem Überblick über Systeme, die eine inhaltliche Erschließung von Freitexten ermöglichen. Die Beschreibung erfolgt entlang von zwei Dimensionen. Zum einen wird der Anwendungskontext der Systeme berücksichtigt (i.e. allgemeine Texte und speziell medizinische Texte) und zum anderen die Verstehensleistung der Systeme (z.B. Klassifikation, Codierung oder Faktenextraktion) in die Betrachtung mit einbezogen.

2.1 Informationsextraktionssysteme

Im Gegensatz zu den NLP-Systemen sind IE-Systeme stärker für die Anwendung mit größeren Datenmengen und zugleich engerem Fokus auf die in Texten enthaltene Information konzipiert. Aus der größeren Anzahl an Vertretern kommen hier die beiden Systeme FASTUS und SCISOR zur Darstellung.

2.1.1 Das FASTUS-System

Mit FASTUS präsentiert sich ein System, das sich mit einer einfachen, auf endlichen Automaten basierenden Architektur ganz der Informationsextraktionsaufgabe verschrieben hat und bewußt von vornherein auf eine tiefere inhaltliche Analyse verzichtet (Appelt

et al., 1993). Damit steht die Abbildung von Texten auf eine begrenzte Menge vordefinierter Informationsmuster (templates) im Vordergrund. Einhergehend mit dem flacheren Textverstehen erreicht das System hohe Verarbeitungsgeschwindigkeiten bei der Informationsextraktion. Die eigentliche Verarbeitung beschränkt sich darauf, Nominalgruppen und Verbalgruppen zunächst einzeln zu analysieren. Die hieraus resultierenden syntaktischen Strukturen werden im nachfolgenden Schritt auf sogenannte *Event-Pattern* abgebildet, die die letzte Verarbeitungsphase schließlich unter Berücksichtigung semantischer Wohlgeformtheitsbedingungen zusammenfaßt. Das syntaktische Modul ist domänenunabhängig konzipiert, während die Abbildung auf die inhaltliche Repräsentation und das Zusammenfassen der Informationsmuster (die Vereinigung von Teilberechnungsergebnissen) an die jeweils unterschiedlichen Anwendungsdomänen anzupassen ist. Aufgrund der schlanken Architektur konnte FASTUS mit minimalem Aufwand an die MUC-4-Aufgabe angepaßt werden, wobei die Werte für Recall und Precision mit 44% bzw. 55% im Blindtest nur von einem anderen System im Testfeld signifikant übertroffen wurden. Spätere Versionen von FASTUS, die sich im Rahmen von MUC-5 und MUC-6 einer Evaluation unterzogen, schalten der Analyse noch einen Präprozessor vor. Dieser hat die Aufgabe, Multiwortlexeme und Eigennamen zu identifizieren, um dadurch Effizienz und Effektivität des Systems zu verbessern (Hobbs et al., 1996).

2.1.2 Das SCISOR-System

Das SCISOR-System (Rau et al., 1989) analysiert automatisch Wirtschaftsnachrichten über Firmenübernahmen bzw. Firmenfusionen, die über einen Online-News-Ticker laufen. Als Ergebnis des Analyseprozesses erzeugt SCISOR eine einzige Inhaltsrepräsentation pro Eingabetext und fügt diese einer zentralen Datenbank hinzu.

Grundsätzlich besteht SCISOR aus zwei Analyseschichten, von denen allerdings nur die zweite für die Informationsextraktion zuständig ist. Zunächst sortiert die erste Anwendungsschicht mit einem fünfstufigen Filter, der aus unterschiedlichen Keyword- und Pattern-Matchern besteht, die über den Nachrichtenticker eingehenden Texte in relevante und irrelevante Dokumenten. Im anschließenden Schritt zur Informationsextraktion werden nur noch die thematisch tatsächlich relevanten Texte weiterverarbeitet (Jacobs & Rau, 1990). Das eigentliche Modul, mit dem die Eingabetexte verarbeitet werden, kombiniert eine Bottom-Up- mit einer Top-Down-Analyse. Der Parser berechnet in einer Bottom-Up-

Strategie partielle syntaktische Strukturen, die als Phrasenstrukturbäume kodiert werden. Die semantische Analyse übersetzt diese Teilbäume in frameartige Strukturen zur Wiedergabe des Inhalts. Kombiniert werden diese Teilergebnisse durch eine top-down operierende, erwartungsbezogene Parsingstrategie, die auf konzeptuellen Templates für Firmenübernahmen bzw. -fusionen aufsetzt. Das endgültige Resultat des Informationsextraktionsprozesses besteht aus einem Frame für einen Firmenübernahme- bzw. Firmenfusionsereignis, der mit aus dem Text zu gewinnenden Informationen angereichert ist.

Neben der Domäne “Wirtschaftsnachrichten” kann SCISOR auch die im MUC-2 Korpus verwendeten kurzen Navy-Nachrichten verarbeiten. Die Adaption an diese Domäne nahm mit nur etwas mehr als einem Personenmonat verhältnismäßig wenig Zeit in Anspruch. Aufgrund der simplen Struktur und der engen Fokussierung auf Informationsfragmente in einem Text kann SCISOR ohne Zweifel als IE-System angesehen werden. Da jedoch der Parser zusammen mit seinem semantischen Interpreter in anspruchsvolleren Anwendungskontexten zur Verwendung kommt – beispielsweise in einem Dialogsystem mit Sprachgenerierung (Jacobs & Rau, 1990) –, ist SCISOR nicht prinzipiell auf IE-Aufgaben beschränkt.

2.2 Textverstehenssysteme

Arbeiten zur Wissensextraktion aus Texten haben in der Computerlinguistik, aber auch in der Medizinischen Informatik eine lange Tradition. Entsprechend groß ist die Anzahl von Systemen, die in eine Darstellung mit einbezogen werden könnten. Weil in dieser Arbeit selbstverständlich kein umfassender Überblick zu der ganzen Palette an Systemen gegeben werden kann, beschränkt sich die Darstellung auf eine kleine Anzahl typischer Vertreter, die im Vergleich zum Textverstehenssystem SYNDIKATE heranzuziehen sind, um besondere Paradigmen und Merkmale zu exemplifizieren. Aufgrund des interdisziplinären Kontexts dieser Arbeit finden hierbei insbesondere auch Vertreter aus der Medizinischen Informatik Berücksichtigung.

Der erste Teil in Kapitel 2.2.1 beschreibt zunächst Textverstehenssysteme, die im Kontext der medizinischen Informatik entstanden sind, beginnend mit dem prominentesten Vertreter, dem Linguistic String Project (LSP; Sager et al. (1987)). Kapitel 2.2.2 geht auf vergleichbare Systeme im computerlinguistischen Bereich ein. Wie für die im späteren Kapitel 5.3 thematisierten Wissensrepositorien gilt auch für die Textverstehenssysteme, daß ein Wissens- und Methodentransfer zwischen den beiden Wissenschaften überraschender-

weise so gut wie nicht stattfindet. Als einige der wenigen Ausnahmen sei daher an dieser Stelle auf das LSP verwiesen, da der Parser von PUNDIT (siehe unten Kapitel 2.2.2.1), einem CL-System, auf einer Prologimplementierung des LSP-Parsers basiert. Umgekehrt verwendet das hier nicht näher thematisierte SPECIALIST-System (McCray et al., 1994) – vereinfacht ausgedrückt ein Folgeprojekt zu LSP – sowohl den Parser als auch die Grammatik von PUNDIT (Spyns, 1996).

2.2.1 Medical Language Processing

Innerhalb der medizinischen Informatik, die aufgrund des breiten Forschungsumfelds aus vielen Forschungs- und Anwendungsgebieten besteht, hat die Analyse von frei verfaßten Texten mit medizinischem Inhalt einen festen Platz, und es existiert eine ausreichende Anzahl von Textverstehenssystemen für die Prägung des Begriffs *Medical Language Processing* (MLP). Im Gegensatz zu den NLP-Systemen im computerlinguistischen Bereich gibt es ein reduziertes Interesse an linguistischer Theoriebildung und ein gesteigertes an verwertbaren Systemen, die die vielfältigen Informationsbedürfnisse in der Medizin zu bedienen in der Lage sind.

Die Anwendungsorientierung drückt sich beispielsweise in einem starken Zwang zur Evaluierung von Verstehensleistungen aus sowie dem Anspruch, mittelfristig auch einsetzbare Systeme zu entwickeln. Sowohl die (positiv zu vermerkende) Ausrichtung auf Anwendbarkeit, als auch die (leider weniger erfreuliche) Tendenz, relativ unabhängig von der computerlinguistischen Community zu arbeiten, rechtfertigen es, die entsprechenden Systeme in einem separaten Unterkapitel abzuhandeln.

2.2.1.1 Der Medical Language Processor des Linguistic String Projects

Obwohl die Zielstruktur der textuellen Analyse des Linguistic String Projects (Sager et al., 1987) aus einer begrenzten Anzahl normalisierter Informationsschablonen besteht, die den Inhalt unterschiedlicher Inhaltssegmente eines Eingabetexts aufnehmen, muß dieses System aufgrund seiner komplexeren Architektur – wie sich etwa durch die semantische Interpretation von Komposita (Wolff, 1984) belegen läßt – zu den NLP-Systemen gezählt werden. Neben dem ursprünglichen englischsprachlichen System gibt es eine französische Version sowie einen deutschsprachlichen Prototypen (Spyns, 1996). Innerhalb der Medizin existieren LSP-basierte MLP-Anwendungen in einer Reihe unterschiedlicher Teildomänen (Sager

et al., 1994). Der Ansatz des Textverstehenssystems ist zwar innerhalb der medizinischen Domäne weitgehend verallgemeinerbar, durch seine extreme Anpassung an die medizinische Fachsprache bleibt er in wesentlichen Teilen auf diese beschränkt. Hauptsächlich konzentriert man sich darauf, typische Kookkurrenzmuster von Wörtern in Texten zu analysieren und die semantischen Bezüge zwischen Wortoberflächen direkt in die Regeln der *sublanguage grammar* hineinzukodieren. Damit entstehen Muster, die in der Medizin typische Hauptaussagen kodieren (z.B. ein SYMPTOM verbunden mit einem Teil des Körpers, konzeptuell als BODY-PART repräsentiert, das entweder in Form eines Adjektivs oder einer Präpositionalphrase das SYMPTOM modifiziert).

Der Analyseprozeß bildet den Eingabetext auf sogenannte *Information Formats* ab, die jeweils kanonisch für einen bestimmten Hauptaussagetypus stehen. PATIENT-STATE ist beispielsweise das in klinischen Dokumenten am häufigsten vorkommende Information Format. Ein PATIENT-STATE hält eine Menge an bezeichnungslosen “Slots” bereit, um inhaltlich festgeschriebene Teilstrings des Eingabetexts aufzunehmen (Sager et al., 1994). Jedoch vermischen diese Slots syntaktische und konzeptuelle Kategorien. Im PATIENT-STATE stehen SUBJECT und VERB neben PROCEDURE oder INFLUENCE.

Im Gegensatz zu den meisten anderen Textverstehenssystemen verwendet das Sprachverarbeitungsmodul des LSP (LSP-MLP) keine formale Wissensrepräsentationssprache, sondern setzt auf einem aus der Datenbankwelt stammenden Entity-Relationship-Modell auf. Der letzte Interpretationsschritt nimmt eine Abbildung der den Textinhalt wiedergebenden instantiierten Information Formats in eine SQL-Datenbank vor. Das Entity-Relationship-Modell dient hierbei dazu, eine begrenzte Anzahl von Relationen zwischen den abzuspeichernden Information Formats einzuziehen², um diese mit weiteren Inferenzen anzureichern.

Die Ergebnisse der Textanalyse in einer SQL-Datenbank abzuspeichern, bietet den großen Vorteil, daß die Retrieval-Schnittstelle über eine wohldefinierte und weitverbreitete Anfragesprache unmittelbar zur Verfügung steht. In den Tests zur Evaluation der Versteheleistung erfolgt entsprechend ein Abgleich der Ergebnisse aus Datenbankabfragen mit Klassifikationen, die von Experten vorgenommen wurden (Sager et al., 1992). Empirische Untersuchungen in dieser Richtung beschränken sich indessen auf recht einfache Versteheleistungen. Jedoch sind die Werte für Recall und Precision dann entsprechend gut. In

²Überdies vermischen diese Relationen konzeptuelle und syntaktische Kategorien. Beispielsweise finden sich in den Information Formats Relationen wie *subject* und *influence* (Sager et al., 1994).

Sager et al. (1994) finden sich zum Beispiel Werte von 92,5% bzw. 98,6% für die Erkennung von 13 Qualitätssicherungskriterien, die bei der Behandlung von Asthma wichtig sind.

Die vom LSP-MLP erzeugten Repräsentationen sind zudem für weitere inferentielle Dienste nutzbar. In diesen Bereich gehört die automatische Zuordnung von SNOMED-Klassifikationscodes zu Texten (Sager et al., 1995).

2.2.1.2 Das MEDLEE-System

Einen vom Ansatz her einfachen und primär an der Effizienz der Verstehensleistung orientierten Ansatz verwendet das MEDLEE-System (Friedman et al., 1995a). Der thematische Fokus des Systems liegt ursprünglich auf Befundberichten, die aus der Radiologie stammen, mit der weiteren Einschränkung auf den Thoraxbereich. Die Analyse der Eingabetexte erfolgt in drei Schritten. Zunächst wird ein Eingabetext geparkt, anschließend werden die Ergebnisse des Parsing normalisiert und schließlich nach der Herstellung eines kanonisierten Formats Datenbankeinträge mit Hilfe eines kontrollierten Vokabulars vorgenommen (Friedman et al., 1994).

Die zugrundeliegende Grammatik, die in Form einer DCG mit ca. 350 Regeln notiert ist, verwendet keine syntaktischen Non-Terminale, sondern kodiert alleine typische semantische Muster in medizinischen Texten wie beispielsweise DEGREE + CHANGE + FINDING, um die hochfrequenten Okkurrenzen dieses Typs analysieren zu können (z.B.: *mildly increased congestion*). In die Regeln sind darüber hinaus eine Reihe von Wortoberflächen fest einkodiert (z.B.: *and, or, at, near, under*), mit deren Hilfe Nominalgruppen koordinierbar sind (Friedman et al., 1994). Obwohl sich innerhalb dieser semantischen Grammatik durch sehr abstrakte Konzepte (wie beispielsweise FINDING) in den Regeln ein gewisser Grad an Abstraktion findet, leidet die Verallgemeinerbarkeit und die Übertragbarkeit unter dieser sehr an Regularitäten innerhalb der Medizin angepaßten Spezifikationsphilosophie. Die Portierung von einer medizinischen Subdomäne auf eine andere scheint indessen bei gleichzeitiger Beibehaltung der Textsorte Befundbericht keine substantielle Erweiterung der Regelmengen zu erfordern (Friedman, 1997).

Die Ergebnisse des semantischen Parsings verarbeitet die Komponente zur Phrasennormalisierung weiter. Diese führt Formulierungsvarianten eines medizinischen Sachverhalts auf eine kanonische Basisform zurück. Als Grundlage dieser Normalisierungen dient eine Menge von im Lexikon abgelegten Einzelwort- und Phraseneinträge, wobei erstere ca. 1.700 Einträge ausmachen und letztere ca. 1.400 Einträge, und eine Menge an geringgradig

abstrahierenden Abbildungsregeln.

Im letzten Interpretationsschritt erfolgt eine Abbildung der inhaltlichen Repräsentation der Röntgenbefunde auf ein kontrolliertes Vokabular im Format des Medical Entry Dictionary (Cimino et al., 1994). Auch dieser letzte Schritt ist regelgestützt und erzielt die kontrollierten Einträge durch weitere Strings substitutionen für Synonyme sowie durch die Füllung vordefinierter kanonischer Konzepte für Radiologiebefunde. Entsprechenden Stellenwert gewinnt der Konsens in der medizinischen Community darüber, wie eine kanonische Repräsentation auszusehen hat (Friedman et al., 1995b). Die kanonisierten inhaltlichen Repräsentationen eines Befundberichts können schließlich in das elektronische Krankenhausinformationssystem übertragen werden (Johnson & Friedman, 1996).

Einen besonderen Stellenwert besitzen die Arbeiten am MEDLEE-System, weil – für Textverstehenssysteme mit einem wissensbasierten Ansatz und breiter semantischer Interpretation recht untypisch – von Anfang an die Verstehensleistung evaluiert wurde und im Zuge dessen eine Anzahl interessanter Arbeiten zur Evaluierungsproblematik entstanden ist (vgl. Kapitel 6.1).

In einer ersten Arbeit wurden 200 Befundberichte aus der Radiologie zufällig ausgewählt. Die Texte mußten dann von zwei Gruppen Spezialisten, bestehend aus Internisten und Radiologen, und einer Laiengruppe klassifiziert werden (Hripcsak et al., 1995). Freilich gestaltete sich die Klassifikationsaufgabe recht einfach. Es war lediglich eine Auswahl zu treffen, welche von insgesamt sechs Symptomen vorlagen. Die automatische Lösung der gleichen Aufgabe wurde MEDLEE wie auch drei anderen, einfacheren computergestützten Verfahren übertragen, um auch für das System eine Vergleichsgruppe zu haben. Im Resultat gab es bei der Klassifikation zwischen den Experten eine hohe Übereinstimmung. MEDLEE erreichte mit einem Recall von 81% und einer Precision von 98% das Niveau des Experten und war deutlich besser als die beiden Vergleichssysteme. Dennoch sind die guten Werte für Recall und Precision aufgrund der auf recht einfache Muster zugespitzten Extraktionsaufgabe zu relativieren.

Zum Abschluß sei darauf hingewiesen, daß MEDLEE in das Klinikinformationssystem integriert ist (Friedman et al., 1995a) und pro Tag im Durchschnitt ca. 640 Radiologieberichte analysiert, codiert und in den Datenbanken des Columbia-Presbyterian Medical Center ablegt (Jain & Friedman, 1997).

2.2.1.3 Das RECIT-System

RECIT benutzt eine einfache und in besonderem Maße an die Medizinsprache angepaßte Strategie zur Berechnung syntaktischer Teilstrukturen, die als *proximity processing* bezeichnet wird. Nominale Strukturen, welche in der Medizin hochfrequent auftreten, dienen als Ansatzpunkt für die Gruppierung von zu den Nomina gehörigen Wörtern im Satz. Dadurch entstehen pro Eingabesatz eine Reihe von strukturellen Inseln, die zunächst in sich, unabhängig vom sie umgebenden Kontext inhaltlich interpretiert werden (Baud et al., 1992). Mit dieser ad-hoc Parsing-Strategie ist RECIT hochgradig an die Besonderheiten der medizinischen Sprache adaptiert, die für ihre telegrammstilartige Nominalsprache bekannt ist. Die grammatischen Regeln vermischen syntaktische und semantische Kategorien, dadurch daß Konzeptsymbole als Non-Terminale fungieren, um semantische Wohlgeformtheitsbedingungen in die Regeln einzuführen. Hierdurch sind die Regeln sehr speziell und die Regelmenge wird ausgeweitet, da in der Grammatik syntaktisch gleiche Strukturen durch die Verwendung unterschiedlicher konzeptueller Symbole vielfach zu spezifizieren sind. Im Resultat führt die Vermischung der Wissens Ebenen dazu, daß bei der Höher-Skalierung des Systems und insbesondere bei der Portierung mangels Abstraktion ein Großteil der Regeln neu zu schreiben ist, um die semantischen Wohlgeformtheitsbedingungen verfügbar zu machen. Darüber hinaus schließt der Fokus auf lokale sprachliche Cluster die Auflösung anaphorischer Phänomene innerhalb des Satzes wie auch zwischen Sätzen oder gar Diskurssegmenten aus.

In einem nächsten Schritt übertragen Transformationsregeln die semanto-syntaktischen Berechnungsergebnisse in eine kanonische Repräsentation, wobei diese Regeln teilweise die – grundsätzlich partiellen – Berechnungsergebnisse zusammenfassen. Auf dieser Ebene dienen konzeptuelle Graphen als Wissensrepräsentationssprache (Baud et al., 1995). Die Kanonisierung der Repräsentation verwendet ontologisches medizinisches Wissen, das der terminologischen Komponente von GALEN entstammt (Rector et al., 1994) und in der Wissensrepräsentationssprache GRail implementiert ist (Rector et al., 1997). Interessanterweise ist die Sprache für die Repräsentation ontologischen Wissens also nicht identisch mit der Zielsprache für die inhaltliche Repräsentation eines Eingabetexts.

2.2.1.4 Das MENELAS-System

Im Rahmen des EU-Programms Artificial Intelligence in Medicine (AIM) entstand im Projekt MENELAS ein gleichnamiges Textverstehenssystem, das sich auf die Textsorte “Arztbriefe” und die medizinische Subdomäne der Kardiologie spezialisiert hat (Zweigenbaum & MENELAS, 1994). Durch den internationalen Projektrahmen setzt MENELAS von Beginn an in der Konzeption auf eine multilinguale Architektur, die die Sprachen Französisch, Englisch und Niederländisch miteinbezieht.

Die sequentielle Architektur von MENELAS verfügt über einen konventionellen Aufbau, der die semantische Interpretation der syntaktischen Analyse nachordnet. Die semantische Interpretation erfolgt mit Hilfe von semantischen Regeln und eines hierarchischen aufgebauten konzeptuellen Domänenmodells. Als Eingabe in die Interpretation dient ein Tripel bestehend aus zwei Konzepten in der Konzept-Hierarchie sowie einer grammatikalischen Relation. Während der Interpretation erfolgt eine Pfadsuche im Konzeptsystem, das in Form eines Konzeptbaums vorliegt (Bouaud et al., 1995), um eine Menge an konzeptuellen Relationen unter Berücksichtigung der semantischen Regeln zu berechnen. In einem weiteren kanonisierenden Interpretationsschritt berechnet MENELAS aus den intermediären Ergebnissen der semantischen Interpretation normalisierte Repräsentationen. Wie das RECIT-System und auch MEDLEE verwendet MENELAS konzeptuelle Graphen als Wissensrepräsentationssprache, um die inhaltliche Repräsentation aufzunehmen.

Der Textverstehensansatz von MENELAS ist stark wissensbasiert, d.h. feinkörniges und umfangreiches Domänenwissen steuert die dem eigentlichen Textverstehen unterlegten Reasoning-Prozesse, sowohl bei der semantischen Interpretation als auch beim Kanonisierungsschritt. Entsprechend gibt es Ansätze, dieses Wissen durch Analyse des sprachlichen Kontexts, in dem Wörter verwendet werden, durch Bestimmung von Wort-Kookkurrenzen, Vergleiche von Wortkontexten sowie Wortclustering entlang semantischer Achsen im Weltwissen zu verfeinern (Nazarenko et al., 1997). Darüber hinaus präsentieren Zweigenbaum et al. (1995a) detaillierte Überlegungen zum Knowledge Engineering, um das automatische Verstehen medizinischer Freitexte zu unterstützen. Da MENELAS als multilinguales System konzipiert ist, sind die Modellierungen, die dem Textverstehen zugrundeliegen, als gemeinsame Repräsentationsebene für analysierte Texte aller Quellsprachen einzelsprachunabhängig.

In Zahlen ausgedrückt verfügt MENELAS im Weltwissen über 1.800 Konzepte und 300

Relationen, dazu gibt es 500 Referenzmodelle für die Kanonisierung der Repräsentation. Es finden sich darüber hinaus 1.000 semantische Regeln bei 150 grammatischen Relationen. Die Ergebnisse der Pfadsuche leiden trotz Suchheuristiken unter unzureichender Restrangierung, so daß zum Beispiel die Verbindung des Konzepttyps ANGIOPLASTY mit SEGMENT-II im syntaktischen Kontext der französischen Präposition *de* 6.063 mögliche Pfade ergibt, die die Wohlgeformtheitsbedingungen der Pfadsuche erfüllen (Bouaud et al., 1996).

Als eines der wenigen Textverstehenssysteme ist MENELAS – zumindest für französischsprachige Texte – evaluiert (Zweigenbaum et al., 1997). Das zur Evaluation verwendete Verfahren vergleicht das Verhalten des Systems bei Faktenanfragen und Wh*-Queries zu Texten mit den Ergebnissen von Experten (also Medizineren) in einer Art Blackbox. Die Verstehensleistung von MENELAS erreichte recht gute Werte, insbesondere wenn man in Betracht zieht, daß auch die zum Experiment herangezogenen Experten durchaus zu unterschiedlichen Ratings kamen. Ein interessantes Ergebnis bei der empirischen Untersuchung der Textverstehensleistung besteht in der Tatsache, das Textverstehenssysteme für bestimmte Klassen von Benutzeranfragen im Vergleich zu Informationsextraktionssystemen besonders gut geeignet sind. Hierzu gehören Anfragen zu komplexeren Faktenaussagen sowie wissensbasierte Anfragen (Zweigenbaum et al., 1995b; 1997).

2.2.2 Textverstehen in computerlinguistischen Systemen

In Abgrenzung zu den IE-Systemen, aber auch als Pendant zu den MLP-Systemen möchte ich in diesem Unterkapitel wenigstens zwei Textverstehenssysteme aus der Computerlinguistik beschreiben. Von den IE-Systemen unterscheiden sie sich durch eine deutlich tiefere Analyse des Eingabetextes, die sich in semantisch reichhaltigeren Zielstrukturen widerspiegelt. Gegenüber den MLP-Systemen weisen sie eine deutlich stärkere Bindung an linguistische Kategorienbildungen und eine breitere formale Fundierung auf.

2.2.2.1 Das KERNEL-System

Da das KERNEL-System (Palmer et al., 1993) im wesentlichen eine Weiterentwicklung des PUNDIT-Systems mit erweiterten Schlußfolgerungsfähigkeiten darstellt (Palmer, 1990), wird hier nur ersteres einer genaueren Betrachtung unterzogen. Das KERNEL-System er-

zeugt mit einem wissensbasierten Ansatz inhaltliche Repräsentationen, die ein breites Spektrum an beim Textverstehen kritischen Phänomenen abdecken. Daher versteht sich KERNEL als ein *broad coverage, general-purpose* System. Insbesondere die Behandlung textueller Phänomene und die Repräsentation temporaler Beziehungen zwischen Events treten in diesem Zusammenhang als besondere Merkmale in den Vordergrund. Als Textgrundlage dienen unter anderem die in MUC-1 und MUC-2 verwendeten Corpora über Nachrichten, die von Marineeinheiten über Sichtungs- und Angriffseignisse abgesetzt werden.

In der Systemarchitektur gibt es eine sequentielle Abfolge der unterschiedlichen Analyseschritte, die für jede der dabei involvierten Wissens Ebenen eine eigene Repräsentation erzeugen. Diese dient dann dem nachfolgenden Modul als Eingabe zur Weiterverarbeitung. Zunächst erfolgt die syntaktische Analyse, die auf der Grundlage einer *Restriction Grammar* (Grishman & Hirschman, 1986) Prädikat-Argument-Strukturen berechnet. Insgesamt vier Module zur semantischen Interpretation setzen auf den syntaktischen Strukturen auf.

Das *clause analysis* Modul verbindet grammatikalische Relationen, die aus der syntaktischen Eingabe stammen, als Argumentstellen des Matrixverbs mit dessen konzeptueller Repräsentation. Der konzeptuelle Rahmen eines Verbs ist mit dem Eintrag einer Wortoberfläche im Lexikon verbunden, der die thematischen Rollen in Form von *lexical conceptual clauses* repräsentiert. Der semantische Interpretationsprozeß wird dabei von sogenannten *mapping rules* unterstützt, die die syntaktischen Prädikate auf konzeptuelle Prädikate abbilden. Konzeptuelle Verträglichkeitsbedingungen liegen als *semantic class restrictions* in Klauselform vor. Im Resultat ist wegen der Vermischung von Wissens Ebenen eine größere Anzahl an Regeln notwendig, um die unterschiedlichen Wohlgeformtheitsbedingungen zu kodieren. Die *mapping rules* spiegeln dabei die auf Fillmore (1968) zurückgehenden semantischen Abhängigkeiten zwischen Kasus und semantischen Argumenten wider.

In der nachfolgenden *noun phrase analysis* geht es darum, die richtigen Modifikatoren mit den nominalen Köpfen zu assoziieren. Gleichzeitig sind die Referenten zu bestimmen, die das *reference resolution* Modul verarbeitet. Da viele Nomen als Ausdrücke angesehen werden, die in der semantischen Interpretation Prädikate über abhängige Argumente beinhalten, werden sie in vielen Fällen wie die ihnen entsprechenden Verben behandelt und greifen auf dieselben lexical conceptual clauses und semantischen Wohlgeformtheitsbedingungen zurück. Jedoch verwenden sie andere mapping rules, weil sie anderen Wortarten angehören.

Im *reference resolution* Modul werden satzübergreifende anaphorische Phänomene mit

Hilfe von Fokuslisten aufgelöst (Dahl, 1986). Diese beinhalten mögliche Antezedenten aus dem Vortext und referenzieren sowohl Diskursobjekte als auch Events. Mit Hilfe dieser möglichen Antezedenten erfolgt die formale Rekonstruktion des Textzusammenhangs. Beispielsweise kann die nominale Anapher “der Angriff” auf ein im Vortext durch ein Verb angezeigtes Ereignis (in der verwendeten Domäne zum Beispiel “der Zerstörer greift das U-Boot mit Wasserbomben an”) verweisen. In der Struktur, die den anaphorischen Ausdruck enthält, werden dann leere Argumentstellen mit den über den Antezedenten zugreifbaren Argumenten gefüllt; die Referenzidentität zwischen im Diskurs explizit enthaltenen Diskursidentitäten drückt ein spezielles *id*-Prädikat aus. Auf der Ebene der Referenz deckt KERNEL Pronomina, textuelle Ellipsen, nominale Referenzen auf Ereignisse sowie implizite Referenten, die im Text nicht vorkommen, ab.

Das letzte Modul zur semantischen Interpretation dient der Berechnung temporaler Bezüge zwischen Events. Einfache Schlußfolgerungsregeln zur Bestimmung zeitlicher Relationen zwischen Sätzen liegen diesem Modul zugrunde, wobei mit Dowty (1986) angenommen wird, daß die Zeit in einem Text per Default von Satz zu Satz progredient ist. Ereignisse haben entweder eine zeitliche Ausdehnung (*period*) oder finden in einem umschriebenen Augenblick (*moment*) statt. Die Beziehung zwischen über Zeitmarken gekennzeichnete Events wird über ein Relationsinventar beschrieben, das die Events relativ zueinander wie auch zu Referenzzeitpunkten und zum Äußerungszeitpunkt positioniert. Die hierbei verwendeten Relationen verfügen über keine spezielle temporale Semantik, sondern sind einfache Prädikate erster Ordnung, wie folgende in Prolognotation gegebenen Beispiele belegen: `precedesP([moment8], speechTime)` oder `includesP([period4], [moment8])`.

2.2.2.2 Das TACITUS-System

Ein auf der Ebene der Verstehensleistung sehr ambitioniertes System ist das am SRI entwickelte TACITUS System (Appelt et al., 1993). Neben dem Anspruch, das in einem Text enthaltene explizite Wissen vollständig und akkurat zu extrahieren, wird außerdem eine möglichst umfassende Rekonstruktion des implizit in einem Text enthaltenen Wissens angestrebt. Entsprechend den ambitionierten Anforderungen benötigt TACITUS auf allen Wissens Ebenen umfangreiche und detaillierte Wissensbestände. Der Umgang mit authentischen Texten erfordert eine robuste Parsingmaschinerie, die eine Reihe unterschiedlicher Strategien anwendet, um mit ungrammatischer Eingabe, systemseitiger Unterspezifikation und Problemen der Berechnungskomplexität zurechtzukommen. Da die verwendeten

(Zeitungs-)Texte im allgemeinen sehr lange Sätze enthalten, benutzt der Parser lexikalisch bzw. grammatikalisch fundierte Präferenzstrategien, damit die Anzahl struktureller Lesarten nicht explodiert (Hobbs & Bear, 1990). Erreicht das System bei der Analyse eines Satzes keine vollständige Überdeckung, dann sorgt ein Wrap-Up-Verfahren dafür, daß möglichst große partielle Ergebnisse berechnet werden (graceful degradation of performance).

Das Ergebnis des Satzanalyseprozesses ist zunächst eine Menge logischer Axiome, die den expliziten Inhalt repräsentieren. Die formale Rekonstruktion impliziter Aussagen verwendet abduktives Schlußfolgern (Hobbs et al., 1993). Dazu fügt das Modul für die pragmatische Interpretation auf der Grundlage von Allgemeinwissen, das im System vorhanden ist, eine minimale Menge logischer Axiome hinzu. Die hinzugefügten Axiome müssen abduktiv aus dem textuellen Kontext und das Allgemeinwissen begründbar sein. Zur Bestimmung, welche Axiome einzuführen sind, belegen Hobbs et al. (1993) Klassen von Axiomen mit Kosten. Auf der Satzebene deckt die formale Rekonstruktion die Beziehung von Nomen in Komposita ab, auf der textuellen Ebene beispielsweise Referenzphänomene oder textuelle Ellipsen.

Die Verstehensleistung von TACITUS ist im Zuge der MUC-3 (Sundheim, 1991) evaluiert worden, wobei der Recall mit 24% bei einer Precision von 40% im Vergleich kein besonders gutes Ergebnis anzeigen, zumal wenn man die schlechten Laufzeiteigenschaften des Systems in Betracht zieht (Hobbs et al., 1992). Genau hierauf gründet sich die Motivation, mit FASTUS (siehe Kapitel 2.1.1) ein an die Informationsextraktion adaptiertes System zu entwickeln.

2.3 Dedizierte Textverstehenssysteme

Unter dedizierten Textverstehenssystemen sind diejenigen Systeme zu verstehen, die zwar ein tieferes Textverstehen anvisieren, die sich jedoch entweder bei der Textanalyse auf bestimmte Sprachmuster beschränken oder deren Fokus bei der Wissensextraktion sehr eng gefaßt ist. In diesem Sinn nehmen sie eine Mittelstellung zwischen den Informationsextraktionssystemen auf der einen Seite und den allgemeiner ausgerichteten Textverstehenssystemen auf der anderen Seite ein. Insbesondere in der medizinischen Domäne ist eine Vielzahl von spezialisierten Ansätzen zur Textanalyse anzutreffen. Eine Legitimation für diese Fokussierung findet sich im konkreten Anwendungsinteresse, etwa der automatischen Entdeckung bestimmter Risikofaktoren (z.B. einer möglichen Tuberkuloseerkrankung an

der Lunge). Um diesen Systemtypus zu charakterisieren, seien einige Beispiele aus der Medizin herausgegriffen.

Einen Ansatz zur semantischen Interpretation räumlicher und partitiver Beziehungen zwischen anatomischen Konzepten, die in Befundberichten über Koronarangiographien vorkommen, verfolgen Bean et al. (1998). Insgesamt 314 sehr speziell auf das Anwendungsproblem zugeschnittene semantische Interpretationsregeln, die an bedeutungsverwandten Verben hängen, berechnen anatomische Beziehungen unter Verwendung eines Inventars von ca. 30 räumlichen und partitiven Relationen. Um die in den Texten enthaltenen Beschreibungen dem semantischen Interpretationsprozeß zuzuführen, finden ein Reihe von Wissensmodulen Verwendung. Zunächst bildet METAMAP im Text enthaltene Terme auf Konzepte des UMLS-Metathesaurus ab (Aronson et al., 1994). Die im semantischen Netzwerk des UMLS (NLM, 1998) verfügbare Information über semantische Typen dient in Verbindung mit dem anatomischen Wissen, das dem Digital Anatomist der University of Washington entnommen ist (Rosse et al., 1998a), dazu, der semantischen Interpretation das notwendige Domänenwissen zur Verfügung zu stellen.³ Die eigentliche konzeptuelle Relationierung, die mit Hilfe spezieller semantischer Regeln erfolgt, hängt indessen nicht von der Struktur des semantischen Netzwerks oder dem anatomischen Modell des Digital Anatomist ab, sondern ist fest mit den Verben verbunden. Beispielsweise weist ein Verb wie *“arise”* auf eine BRANCH-OF-Relation hin.

In einem Vergleich zur Klassifikationsleistung zweier Annotierer (aus der Autorengruppe) konnte für die semantische Interpretation anhand von 15 randomisiert ausgewählten Befundtexten ein Recall von 78% bei einer Precision von 67% ermittelt werden. Jedoch darf bei einer Bewertung dieser Werte nicht unberücksichtigt bleiben, daß nur ein stark eingeschränkter Teil der Textaussagen überhaupt interpretiert wurde und dies bei einer Regelmenge von 314 Regeln für eine relativ kleine Domäne.

Ein weiteres mögliches Einsatzfeld von Textverstehenssystemen mit eingeschränktem Fokus liegt im Bereich von Decision-Support-Systemen, da ein Großteil an handlungs- und entscheidungsrelevanten klinischen Informationen über Patienten in Freitexten verborgen ist. Um diesem Problem Abhilfe zu verschaffen, gibt es prinzipiell zwei Möglichkeiten. Einerseits können Daten bereits bei der Erstellung erschlossen werden, indem strukturierte Eingabemasken zum Einsatz kommen. Andererseits können Textverstehenssysteme

³Beide Wissensrepositorien werden in Kapitel 5.3.2 genauer beschrieben.

helfen, den Inhalt von Freitexten zugänglich zu machen. Im Kontext von Decision-Support-Systemen sind i.a. nur Wissensfragmente aus den Eingabetexten zu extrahieren.

Jain et al. (1996) stellen ein System vor, das aus radiologischen Befundberichten Informationen gewinnt, die den Verdacht auf Tuberkulose bei einem Patienten anzeigen. Die betroffenen Patienten werden dann entsprechend besonderer Versorgungsrichtlinien behandelt. Damit läßt sich sowohl die Versorgung der Patienten verbessern als auch eine Reduzierung der Kosten herbeiführen. Das MEDLEE System (vgl. Kapitel 2.2.1.2) führt die inhaltliche Analyse der Eingabetexte durch und erzeugt inhaltliche Beschreibungen bezüglich der beurteilungskritischen Merkmale, anhand derer das Vorliegen einer Tuberkulose bestimmbar ist. Im positiven Fall generieren Regeln, die in Form von Medical Logic Modules (Hripcsak, 1994) vorliegen, im System Ausgaben, die die Behandelnden über den Tuberkuloseverdacht informieren.

In der Evaluation gegenüber einem von intellektuellen Klassifikatoren erstellten Goldstandard brachte das System eine Übereinstimmung von nahezu 90% zustande. Berücksichtigt man die Tatsache, daß die Rater untereinander selbst nicht zu 100% übereinstimmen, kann mit dem System die Entdeckung eines Tuberkuloseverdachts durchaus automatisiert werden.

Eine weitere Anwendungsstudie von Jain & Friedman (1997) fällt in den Bereich der automatischen Beurteilung von Mammographien, wobei wieder das Textverstehenssystem MEDLEE verwendet wurde. Da Mammographien ein Bestandteil der Krebsvorsorge sind, fallen im klinischen Alltag große Mengen an Daten in Form von freitextlichen Befunden an. Aus diesem Grund hat die computergestützte Erkennung von auffälligen Befunden einen unmittelbaren Nutzen im Klinikalltag. Ein – allerdings im wesentlichen informell geführter – Vergleich der Verstehensleistung von MEDLEE zu manuellen Auswertungen ergab eine weitgehende Gleichwertigkeit zwischen dem automatischen und dem intellektuellen Verfahren.

Zum Schluß möchte ich noch zwei weitere Anwendungsfelder kurz anführen: Textverstehen, um computergestütztes Lernen zu vereinfachen, und Textverstehen mit dem Zweck, Retrievalergebnisse zu verbessern.

Um Passagen von Lehrbuchtexten mit computergespeicherten anatomischen Darstellungen zu verbinden, führen Sneiderman et al. (1998) eine flache inhaltliche Analyse durch, die das SPECIALIST-System nutzt (McCray et al., 1994). Dabei werden die in den Texten enthaltenen anatomischen Terme auf UMLS-Identifizier abgebildet (NLM, 1998). Im Anschluß

an die syntaktische Analyse folgt die semantische Interpretation, die die relationalen Bezüge zwischen den identifizierten UMLS-Konzepten berechnet.

Jedoch ist eine semantische Analyse, die auf in UMLS kodierte Relationen zwischen semantischen Typen aufbaut, deshalb problematisch, weil dieser Teil der Ontologie vergleichsweise schwach entwickelt ist und zudem logische Inkonsistenzen aufweist (Schulz & Hahn, 2000). Daher greifen die Autoren auf ein eigenes Beschreibungssystem für konzeptuelle Relationen zurück. Im Resultat kann von den Textpassagen im Lehrbuch über entsprechende Hyperlinks auf anatomische Darstellungen zugegriffen werden.

Auch die letzte in diesem Kapitel dargestellte Anwendung konzentriert sich auf eine sehr spezielle Aufgabenstellung, nämlich auf die Suche nach relevanten Artikeln zu speziellen Befunden in einer Datenbank. Das Auffinden von Satzfragmenten mit Befundaussagen in Abstracts, die zu Zeitschriftenartikeln in der MEDLINE-Datenbank abgelegt sind, wird dabei ebenfalls mit flachen NLP-Methoden angegangen (Sneiderman et al., 1996). Die dedizierte Suche nach Befunden soll den gezielten Zugriff auf relevante Artikel über Erkrankungen verbessern. Verglichen mit der manuellen Annotation von Befundaussagen erreicht das automatische Verfahren, das in einem regelbasierten Ansatz semantische Typen aus dem UMLS Semantic Network verwendet, einen Recall von 82% bei einer Precision von 65%.

2.3.1 Zusammenfassung zu Textverstehenssystemen

Vergleicht man die hier kurz dargestellten IE-, MLP- und NLP-Systeme und versucht zugleich deren Gemeinsamkeiten festzumachen, dann lassen sich im wesentlichen drei Gesichtspunkte herausstellen.

Zunächst ist allen Systemen gemeinsam, daß sie nicht nur versuchen, syntaktische Beschreibungen eines Texts zu erzeugen, sondern auch eine inhaltliche Repräsentation aufzubauen. Dabei verarbeiten sie authentische Texte, also Texte, die nicht in irgendeiner Form manuell für die Verarbeitung aufbereitet worden sind.

Bei IE-Systemen ist der thematische Fokus von vorneherein eingeschränkt: aus Textteilen extrahierbare Informationsfragmente sind auf prädefinierte Templates abzubilden, die die wesentliche Informationsstruktur eines eingeschränkten Diskursbereichs enthalten sollen. Demgegenüber sind die MLP-Systeme und die NLP-Systeme vom Ansatz her weiter gefaßt, denn sie versuchen zum Teil deutlich tiefere inhaltliche Repräsentationen aus den

Eingabetexten zu errechnen und verfügen deshalb i.a. auch über detaillierteres Domänenwissen zur Wissensextraktion. Auffällig bei den MLP-Systemen ist die Präferenz für semantische Kookkurrenzmuster, die sich in der medizinischen Fachsprache klar bestimmen lassen. Eine Zwischenstellung nehmen die dedizierten Systeme ein, weil sie – wie IE-Systeme – nur auf sehr präzise gefaßte Wissenspartikel hin Texte analysieren, die dahinterliegende Analysemaschinerie entstammt dabei jedoch NLP-Systemen.

Eine Eigenschaft, die die meisten Systeme teilen, betrifft die semantische Tiefe der erzielten Verstehensleistung. Im allgemeinen benutzen die meisten Systeme eine frameverwandte Zielstruktur. In den Texten enthaltene essentielle semantische Informationen (z.B. Quantifikation, temporale und spatiale Bezüge, Beziehungen zwischen Individuen und Klassen) werden i.a. nicht adäquat behandelt, sondern entweder ignoriert oder mit unzulänglichen Mitteln angegangen. Dies resultiert aus den Komplexitätseigenschaften und Ressourcenbeschränkungen, die in realistischen Szenarien unmittelbar auf die Verstehensleistungen zurückwirken.

Das nachfolgende Kapitel präsentiert ausführlich das Textverstehenssystem SYNDIKATE, das in die Kategorie der NLP-Systeme fällt, welches aber gleichermaßen medizinische Texte zu analysieren in der Lage ist.

Kapitel 3

Das Textverstehenssystem

SYNDIKATE

Das Textverstehenssystem SYNDIKATE wird seit 1991 in der Arbeitsgruppe Computerlinguistik an der Universität Freiburg im Breisgau entwickelt. SYNDIKATE ist ein Akronym für SYNthesis of Distributed Knowledge Acquired from TEXTs. Die erste Version von SYNDIKATE war auf die Domäne der Informationstechnologie (IT) beschränkt. Mit dem Jahr 1996 begann die Erweiterung auf eine Teildomäne der Medizin (MED), i.e. die der gastrointestinalen Erkrankungen. Die dadurch entstandenen Spezialisierungen des allgemeinen Textverstehenssystems tragen die Bezeichnung ITSYNDIKATE für die Informationstechnologie und entsprechend MEDSYNDIKATE für die Medizin. Die Erweiterung der ersten Version von SYNDIKATE um eine weitere Domäne, deren ontologische Übereinstimmung mit der Informationstechnologie sehr gering ist, führte im Resultat zur Extraktion eines domänenunabhängigen Kernsystems. Abb. 3.1 verdeutlicht den zugrundeliegenden Zusammenhang. Eine übliche Möglichkeit, ein existierendes System an eine andere Domäne anzupassen, besteht in einer vollständigen Portierung (Abb. 3.1 a.). Hierbei wird eine Kopie des Ausgangssystems zur Grundlage des Zielsystems. Dieses Neusystem wird dann durch die entsprechenden speziellen Adaptionen und Erweiterungen der neuen Domäne angepaßt. Abb. 3.1 b. zeigt indessen den in SYNDIKATE beschrittenen Weg der Extraktion eines Kernsystems. In dieses Kernsystem können Erweiterungen durch beliebige Subdomänen mittels Spezialisierung integriert werden, ohne daß tiefergehende Modifikationen am Kern erforderlich werden. Die generelle Erweiterbarkeit von SYNDIKATE ist zurückzuführen auf den Grad an Abstraktion und Allgemeinheit, der für die Teilkomponenten des Textverste-

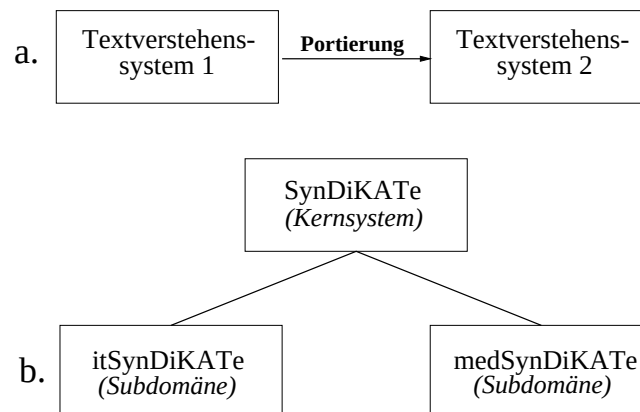


Abbildung 3.1: Modelle der Adaption an Subdomänen

henssystems von Anfang an durch die Konzeption angestrebt war. Im nachfolgenden Text gelten die folgenden Benamungskonventionen: SYNDIKATE bezeichnet den domänenunabhängigen Kern des Textverstehenssystems. Die domänenabhängigen Spezialisierungen sind jeweils durch ein Präfix gekennzeichnet (IT bzw. MED).

Bedingt durch die besonderen Anforderungen, die die Wissensakquisition aus authentischen Texten zweier heterogener Diskursbereiche (i.e., Informationstechnologie und Medizin) mit sich bringt, folgt SYNDIKATE einer Reihe von Entwurfs- und Designprinzipien, welche in diesem Kapitel erläutert und tiefer motiviert werden sollen.

- Die Verarbeitung authentischer Texte erfordert die **Robustheit** der syntaktischen Analyse gegenüber Spezifikationslücken, ungrammatischer Eingabe sowie parametrischem Jargon.
- Die **Lokalität** der syntaktischen und semantischen Berechnungen erlaubt die Erzeugung von Teilergebnissen aus einem Eingabetext und macht das System robust, da keine vollständige analytische Überdeckung des Eingabetexts erforderlich ist.
- Die **Inkrementalität** des Parsing bringt eine wortweise Analyse des Eingabetextes von links nach rechts in Analogie zum menschlichen Sprachverstehen mit sich, ohne daß explizit der Anspruch der kognitiven Adäquatheit der Sprachanalyse erhoben wird. Damit unterstützt das Textverstehenssystem den inkrementellen Aufbau von Bedeutungsstrukturen aus der Eingabe (Hahn & Romacker, 1999a). Soweit es die

bisher erreichte Analyse erlaubt, werden dabei alle möglichen syntaktischen und semantischen Berechnungen nach der Konsumierung eines neuen Wortes abgeschlossen.

- **Kompositionalität** der semantischen Interpretation. Ausgehend von der inkrementellen Analyse bauen sich die Ergebnisse einer inhaltlichen Interpretation des Eingabetextes kompositional aus der Kombination der jeweils berechneten Teilergebnisse auf (Montague, 1974). Die gewonnene inhaltliche Repräsentation reflektiert die erzielte Tiefe des Textverstehens, d.h. den Grad an syntaktischer Überdeckung und erreichter semantischer Interpretation.
- Die konsequente **Modularisierung** der Wissenskomponenten. Unter Modularisierung ist zu verstehen, daß die Wissensquellen weitgehend unabhängig voneinander ausgearbeitet und implementiert werden. Sie kommunizieren lediglich über wohldefinierte Schnittstellen miteinander. Die strenge Modularisierung der Wissens- und Schlußfolgerungskomponenten ermöglicht zum einen die Flexibilität hinsichtlich der Anwendungsdomänen dadurch, daß in Teilmodulen notwendige Adaptionen nicht interferieren.¹ Darüber hinaus ist sie Voraussetzung dafür, daß ein größeres Projekt wie SYNDIKATE organisatorisch abwickelbar bleibt.
- Die **Abstraktion** der Beschreibungen durch Objektorientierung und Vererbung ermöglicht deren sparsame Kodierung, wodurch die Wartbarkeit, die Skalierbarkeit und die Portierbarkeit des Systems unmittelbar unterstützt wird.

Im nachfolgenden Überblick werden die Komponenten des Textverstehenssystems kurz erläutert, und an einem Beispielsatz wird aufgezeigt, wie diese während der Analyse zusammenspielen. Kap. 3.2.1 behandelt die Grammatik, den Parser und das Lexikon, Kap. 3.2.2 das Domänenwissen und bietet einen ersten oberflächlichen Einstieg in die Semantik. Das Kap. 3.2.3 geht auf das Versionierungs- und Konzeptlernen ein, welche für die Aktualität und die Erweiterung des verfügbaren Domänenwissens sorgen. Nach der Einführung der wesentlichen Wissensquellen verdeutlicht das Kapitel 3.2.4 deren Zusammenspiel bei

¹Beispielsweise bringt die Vermengung konzeptuellen und grammatischen Wissens in Semantischen Grammatiken mit sich (z.B. Friedman et al. (1995a)), daß die Einführung neuer Konzeptbeschreibungen (Ebene des Domänenwissens) die Ausweitung der Regeln zur Syntaxanalyse (Ebene syntaktischer Beschreibungen) hervorruft (vgl. hierzu die Ausführungen in Kapitel 2.2.1).

der Satzanalyse. Schließlich beschäftigt sich Kap. 3.2.5 mit den Systemerweiterungen zur Behandlungen von Textphänomenen.

3.1 Anwendungskontext von SYNDIKATE

SYNDIKATE verarbeitet authentische Texte, das heißt Texte, die ohne weitere redaktionelle Überarbeitung den benutzten Textkorpora entnommen werden. Bei den verwendeten Texten handelt es sich zum einen um Artikel zu verschiedenen Themenbereichen aus einschlägigen deutschsprachigen Computerzeitschriften mit unterschiedlichen fachlichen Anforderungen an die Leser, zum anderen um routinemäßig im Klinikalltag erstellte Befundberichte aus dem Bereich Histopathologie bei gastrointestinalen Erkrankungen. Letztere entstammen dem Klinik-Informationssystem der Universitätsklinik Freiburg (Klar et al., 1991). Durch dieses realistische Anwendungsszenario ergeben sich folgende Problemfälle bezüglich der Eingabe in das Textverstehenssystem:

- *Ungrammatische Eingabe*: Die Eingabe entspricht nicht den Regeln einer Standardgrammatik. In diesen Bereich gehören beispielsweise Kongruenzfehler, Interpunktionsfehler, Orthographiefehler etc.
- *Paragrammatische Eingabe*: Die Eingabe entspricht nicht den Regeln einer Standardgrammatik, ist jedoch für Subgruppen einer Sprachgemeinschaft akzeptabel. Insbesondere in Routinetexten aus der Medizin spielt dieses unter dem Label “Sublanguage” bekannte Phänomen eine größere Rolle (Dunham, 1986). In der Medizin lassen sich paragrammatische Bildungen sowohl auf der Wortebene als auch auf der Satzebene beobachten (vgl. Heckl (1990)). Die enorme Produktivität der medizinischen Sprache bei der Komposition und Derivation (Wortebene) erlaubt unübliche Ableitungen wie z.B. *intraepithelial* als Ortsbezeichnung oder *lymphoplasmazellulär* als adverbiale Bestimmung des Agenten einer Handlung.² Daneben ist der als Telegrammstil bekannte Jargon der Mediziner in Texten wie Arztbriefen, Befundberichten und Entlassungsberichten das Standardbeispiel für den paragrammatischen Stil dieser soziolektalen Sprachgemeinschaft (Satzebene).

²Im allgemeinen entstammen die hier verwendeten Beispiele den Textkorpora. Sind Beispiele konstruiert, so wird dieses gesondert vermerkt.

- *Extragrammatische Eingabe*: Die Eingabe entspricht zwar den Wohlgeformtheitsbedingungen einer Standardgrammatik, jedoch kann das Textverstehenssystem die Eingabe nicht vollständig analysieren, weil Spezifikationslücken bestehen. Diese Spezifikationslücken können verschiedene Ebenen betreffen. Zum Beispiel deckt die im System vorhandene Grammatikspezifikation im Normalfall nur einen Teil der Grammatik einer natürlichen Sprache ab.³ Auch ergeben sich Abdeckungsprobleme auf der Ebene des Lexikons durch Wortkomposition und Wortderivation.

Das Problem unvollständiger sprachlicher Wissensquellen in Verbindung mit ungrammatischer beziehungsweise paragrammatischer Eingabe stellt sich für jede Form der Analyse authentischer Sprache. Das heißt, das Analysesystem muß auch unter diesen realistischen Rahmenbedingungen in der Lage sein, ein Analyseergebnis zu liefern, ohne in eine Analyseblockade zu geraten. Genau dies umschreibt das informelle Kriterium für *robustes Parsing*. Im besten Fall nimmt die Qualität der Analyse beim partiellen Textverstehen proportional zum Grad der Unterspezifikation ab (*graceful degradation*).

3.2 Architektur von SYNDIKATE: Überblick

Als Eingabe in das Textverstehenssystem SYNDIKATE dient eine Kollektion von Texten, die unseren Korpora für Informationstechnologie und Medizin entstammen (vgl. Abb.3.2 oben). Der Textverstehenskern transformiert jeden Eingabetext in eine entsprechende Textwissensbasis, die den Textinhalt in einer formal fundierten Wissensrepräsentationssprache enthält (siehe Abb.3.2, Text KBs). Anschließend an diese Phase der automatischen Textwissensakquisition kann das in den Textwissensbasen formal kodierte Wissen für vielfältige Informationsdienste genutzt werden, was in der Abb.3.2 durch die Synthese der Textwissensbasis und der Bereitstellung eines Wissenspools grafisch angedeutet wird. In der Medizin zählen dazu etwa die Befundverschlüsselung nach den aktuellen Verschlüsselungsrichtlinien (entsprechend den gegenwärtigen gesetzlichen Bestimmungen erfolgt in Deutschland die Diagnosenverschlüsselung beispielsweise nach ICD-10 (1992)). Im weiteren Horizont

³Nicht behandelt wird an dieser Stelle das Problem, inwieweit die in Textverstehenssystemen verwendeten Grammatiken eine adäquate Ausdrucksmächtigkeit besitzen, um alle Phänomene einer natürlichen Sprache formal zu fassen. Zum Beispiel sind die häufig anzutreffenden kontextfreien Grammatiken nicht in der Lage, bestimmte Cross-Over Phänomene im Züricher Deutsch (Shieber, 1985) oder im Niederländischen (Rentier, 1994) zu kodieren.

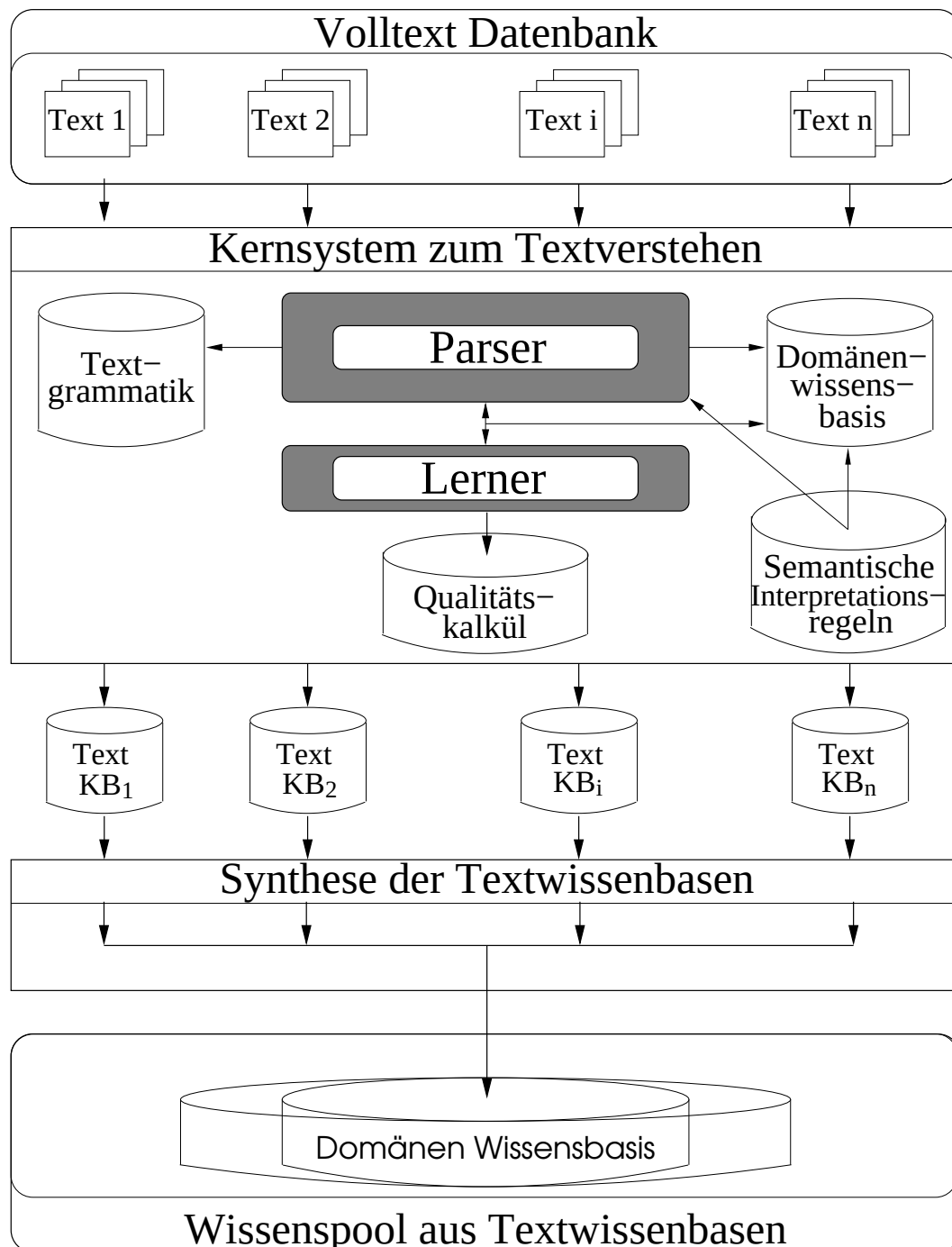


Abbildung 3.2: Interaktion der Wissensmodule im Textverstehenssystems SYNDIKATE

des Textverstehens, der sich nicht nur auf die Medizin beschränkt, gehören dazu ferner die Textklassifikation, die Textfiltrierung, das Textpassagen-Retrieval sowie das inferentiell gestützte Faktenretrieval, aber auch sehr wissensintensive Dienste wie die kaskadierte Textzusammenfassung (Hahn et al., 1990). Realisiert ist beim augenblicklichen Stand der Implementation allerdings nur das Kernsystem zum Textverstehen (vgl. Abb. 3.2 Mitte), alle weitergehenden Informationsdienste befinden sich noch im Stadium der prototypischen Entwicklung. Abbildung 3.2 zeigt die wesentlichen einzelnen Komponenten von SYNDIKATE, nämlich den Parser (einschließlich Grammatik), die Textgrammatik, das Domänenwissen, die semantischen Regeln sowie den Lerner mit seinem Qualitätskalkül, verdeutlicht deren Zusammenhang und stellt den Kontrollfluß zwischen ihnen dar. Aufgrund des augenblicklichen Stands der Implementation beschränken sich die nachfolgenden Kapitel 3.2.1 bis 3.2.5 auf die Beschreibung der unterschiedlichen Wissensquellen, welche im Kernsystem realisiert sind, sowie deren Interaktion während des Analyseprozesses für einen Eingabetext.

3.2.1 Grammatik, Parser und Lexikon

Auf der Beschreibungsebene liegt das grammatikalische Wissen von SYNDIKATE in Form einer Dependenzgrammatik vor (Bröker et al., 1994).⁴ Dependenzgrammatiken gestatten die Beschreibung von lokalen syntaktischen Wohlgeformtheitsbedingungen zwischen einem syntaktischen Kopf und einem Modifikator, die an eine spezifische Dependenzrelation geknüpft sind. Dazu definieren wir die Menge $\mathcal{D}$ als die Menge aller Dependenzrelationen: $\mathcal{D} := \{\text{subject}, \text{direct-object}, \text{indirect-object}, \text{genitive-attribute}, \dots\}$.⁵ Eine syntaktische Leerstelle an einem lexikalischen Kopf, die von einem Modifikator besetzt werden kann, bezeichnet man als Valenz. Entsprechend ist die Menge $\mathcal{V} := \{\text{AdverbVal}, \text{PrepositionVal}, \text{SubstantiveVal}, \dots\}$ als die Menge aller Valenzen definiert. Zu den lokalen Constraints für eine Valenz, die verschiedene heterogene Wissens Ebenen betreffen und im Begriff der Dependenzrelation zusammengefaßt sind, gehören insbesondere:

⁴In dieser Arbeit werden Entwurfs- und Implementationsprinzipien nur dann motiviert, wenn sie in unmittelbarem Zusammenhang mit dem Thema stehen. Ansonsten erfolgt eine neutrale Beschreibung der Systemumgebung, in der die semantischen Interpretationen stattfinden.

⁵In den Abbildungen zu Dependenzgraphen für Beispielsätze kommen zuweilen Abkürzungen für Dependenzrelationen vor. Im Text finden sich die Bezeichner für Dependenzrelationen indessen grundsätzlich ausgeschrieben.

- die Unifizierbarkeit morphosyntaktischer Merkmalsstrukturen aus der Menge 2^U (dabei sei U die Menge der Merkmalsstrukturen) von Kopf und Modifikator (vgl. hierzu z.B. Shieber (1986)),
- Wortart-Informationen, d.h. Informationen darüber, welcher Wortart ein potentieller Modifikator für eine Dependenzrelation angehören muß (Bröker, 1999),
- Stellungsrestriktionen für einen Modifikator, sogenannte Wortstellungsdomänen, die definieren, welche Position ein in einer bestimmten Wortklasse zugehöriger Modifikator gegenüber seinem syntaktischen Kopf einnehmen kann (Bröker, 1998) und schließlich
- semantisch-konzeptuelle Wohlgeformtheitsbedingungen, die syntaktisch zulässige, semantisch-konzeptuell jedoch zu verwerfende Dependenzbeziehungen ausfiltern und auf diese Weise die Anzahl struktureller Ambiguitäten reduzieren.

Auf der Ebene der Grammatikbeschreibung und des Parsingverfahrens folgt SYNDIKATE zunächst konventionellen Ansätzen. Jedoch bringt die Forderung nach Robustheit des Parsingverfahrens, die die Voraussetzung für die Analyse von authentischen Texte darstellt, ein entscheidendes Re-Design der Grammatik im Sinne einer *Performanzgrammatik* mit sich (Bröker et al., 1996). Innerhalb der Performanzgrammatik sind neben deklarativen Beschreibungen auch *prozedurale* Beschreibungen, die das Verhalten des Parsers (etwa bei der Etablierung von strukturellen und referentiellen Bezügen) bestimmen, Bestandteil des grammatischen Wissens (die formale Charakterisierung dieser prozeduralen Elemente in der Grammatik findet sich in Schacht (1999)).

Naheliegend ist bei dieser einheitlichen Sicht auf Daten (Grammatik) und Prozeduren (Parsing) eine objektorientierte Spezifikation der Grammatik und eine ebenfalls objektorientierte Implementation des Parsers. Ausgehend vom Aktoren-Berechnungsmodell (Agha et al., 1997) erfolgt die objektorientierte linguistische Spezifikation im Rahmen der Performanzgrammatik auf der Grundlage eines um terminologisches Schließen erweiterten Objektkalküls (Schacht & Hahn, 1997). Diese Spezifikationen werden in ACTTALK (Briot, 1989) implementiert, einer Variante der objektorientierten Programmiersprache SMALLTALK (Goldberg & Robson, 1983; 1989), die zusätzlich asynchrone Kommunikation zwischen Objekten zur Verfügung stellt.

Die Verteilung von Berechnungsprozessen auf Aktoren erlaubt die Erzeugung von Teilergebnissen, die als partielle Lösungen eines komplexen Berechnungsproblems weiterverar-

beitet werden können (Schacht et al., 1994). Zusammen mit weiteren Vorzügen des objektorientierten Paradigmas, etwa der Vererbung in Objekthierarchien und der Kapselung von Berechnungsprozessen, bietet dieser Modellierungskontext somit einige inhärente Möglichkeiten, die Lücken in der Spezifikation ohne Blockade der Analyse zu umgehen. Dies unterstützt die Robustheit der Analyse unmittelbar.

Im PARSETALK-System (Bröker et al., 1996), der in SYNDIKATE integrierten Grammatik- und Parsing-Komponente, orientiert sich die Grammatikkonzeption durch eine konsequente lexikalische Verteilung grammatikalischen Wissens ebenfalls am objektorientierten Paradigma. Diese stringente Lexikalisierung erlaubt es auch, daß mittels grammatikalischer Spezifikationen Vorkehrungen für eine nur partielle Analyse von Texten getroffen werden können. Daher laufen im Kontext der verwendeten Dependenzgrammatik Verträglichkeitsberechnungen streng lokal zwischen einem lexikalischen Kopf und einem potentiellen Modifikator ab. Um eine Dependenzrelation zu etablieren, ist - im Gegensatz zu Phrasenstrukturgrammatiken - auf Phrasenebene keine vollständige Überdeckung aller lexikalischen Einheiten notwendig.

Eine Erweiterung des performanzgrammatischen Ansatzes erfolgt durch die Verwendung besonderer Parsing-Protokolle (Neuhaus & Hahn, 1996a) sowie durch die Bereitstellung bewährten verarbeitungsstrategischen Wissens (Hahn et al., 1997) als Teil der lexikalischen Grammatikspezifikation für das *partielle* Parsing, das *prädiktive* Parsing und das *präferentielle* Parsing.

- Mithilfe des partiellen Parsings können Lücken in der Spezifikation auf der Ebene der Lexikoneinträge und der Grammatik überbrückt werden, ohne daß die Analyse des Textes blockiert.
- Das prädiktive Parsing basiert auf der erwartungsbezogenen Satzanalyse und erzeugt im inkrementellen Parsingansatz Voraussagen lexikalischer Köpfe für den rechten Kontext des aktuellen Wortes. Damit können strukturelle Ambiguitäten frühzeitig unterdrückt werden.
- Auf der Grundlage des präferentiellen Parsings selektiert der Parser linguistisch plausiblere Teilergebnisse zur weiteren Analyse, während weniger plausible Kandidaten zunächst nicht weiterverfolgt werden.

Trotz des angeführten verarbeitungsstrategischen Wissens wächst der Suchraum während der Satzanalysen aufgrund lokaler Ambiguitäten stark an. Um im Textverstehenssystem noch ein akzeptables Laufzeitverhalten des Parsers zu gewährleisten, führt der Parser nur ein restringiertes Backtracking aus, bei dem der Suchbaum nicht vollständig abgearbeitet wird. Folglich werden nicht alle möglichen syntaktischen Strukturen berechnet. Das heißt, das Parsingverfahren nimmt bewußt einen Verlust an Strukturbeschreibungen in Kauf. In einer empirischen Untersuchung, innerhalb derer der restringierte Parser aus SYNDIKATE mit einem unrestringierten Chartparser verglichen wurde, ergab sich, daß die Verlustquote gegenüber dem Chartparser bei nur 10 % liegt. Jedoch brachte dies dafür ein lineares Laufzeitverhalten und eine etwa um den Faktor 10 verbesserte Performanz mit sich (Neuhaus & Hahn, 1996b). Darüber hinaus ist bei einer authentischen Eingabe ohnehin nur im Ausnahmefall mit einer vollständig überdeckenden Analyse zu rechnen, so daß die Verlustquote vertretbar erscheint.

SYNDIKATE benutzt beim gegenwärtigen Stand der Implementation ein Vollformenlexikon, d.h., es findet keine Deflexion, Dederivation oder Dekomposition statt. Das Lexikon besteht aus mehreren Teillexika, die entsprechend den inhaltlichen Anforderungen der zu analysierenden Texte zusammengefügt werden können, und folgt damit der Grundidee, Kernsystem und Spezialisierungen auseinanderzuhalten. Das domänenunabhängige Lexikon setzt sich zum einen aus Sublexika für die geschlossenen Wortklassen, zum anderen aus domänenunabhängigen Wörtern der offenen Wortklassen zusammen. Hinzu kommen die Speziallexika für Medizin und Informationstechnologie, die die domänenspezifischen Wörter der offenen Wortklassen vorhalten.

Weiterhin existiert eine Unterteilung der Lexika bezüglich Grund- und Oberflächenform in Lexem- bzw. Wortklassenlexikon. Motivieren läßt sich diese Unterteilung durch die heterogenen Kategorisierungsmechanismen zur Wortbeschreibung, die sich nicht in einer einzigen Beschreibungsebene bündeln lassen, weshalb Wortklasse und Lexemklasse als unabhängige Eigenschaften von Wörtern aufgefaßt werden (Bröker, 1999). So beschreibt das Lexemlexikon den Bedeutungskern eines Wortes sowie dessen Argumentstruktur, d.h. die lexemspezifischen Modifikatoren. Darüber hinaus enthält es den Bezeichner für die mit einem Lexem assoziierten Konzepte im Weltwissen. Im Wortklassenlexikon finden sich hingegen Unterscheidungen bezüglich der syntaktischen Funktion eines Wortes sowie bezüglich dessen morphologischer Flexionsmerkmale und semantischer Eigenschaften (wie Zählbarkeit bei Nomina).

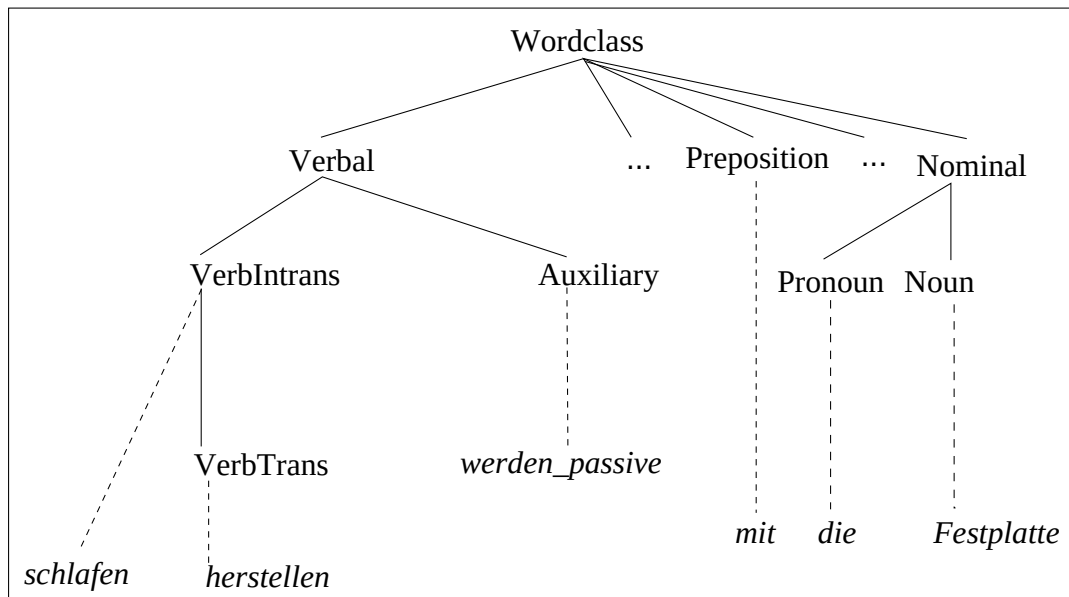


Abbildung 3.3: Wortklassenhierarchie mit lexikalischen Knoten als Blättern

Um innerhalb der lexikalisierten Grammatik die komplexen lexikalischen Beschreibungen ökonomisch zu handhaben, sind diese in Form von Klassenbeschreibungen, die in einer Vererbungshierarchie stehen, realisiert. Dazu definieren wir die Menge der Wortklassen $\mathcal{W}$ mit den Wortklassenbezeichnern als Elementen: $\mathcal{W} := \{\text{VERBAL}, \text{NOMINAL}, \text{NOUN}, \text{PRONOUN}, \dots\}$ sowie die transitive und reflexive Subsumptionsbeziehung $isa_{\mathcal{W}} = \{(\text{VERBINTRANS}, \text{VERBAL}), (\text{ARTICLE}, \text{DET}), \dots\} \subset \mathcal{W} \times \mathcal{W}$. In Analogie hierzu ist die Menge $\mathcal{L}$ für die Lexemklassen mit der transitiven und reflexiven Subsumptionsbeziehung $isa_{\mathcal{L}}$ definiert. Die Wörter selbst bilden die Blätter dieses Baumes und erben die Eigenschaften der Elternknoten.

Eine grafische Darstellung dieses Zusammenhangs für Wortklassen ist in Abb. 3.3 gegeben. Wörter sind in Kursivschrift dargestellt und werden über eine gestrichelte Kante ihren Klassen zugeordnet, während Klassenbeziehungen als durchgezogene Kante gezeichnet sind. Subsumptionsbeziehungen bestehen sowohl zwischen Klassen (z.B. VERBINTRANS für intransitive Verben und VERBTRANS für transitive Verben), als auch zwischen Klassen und Wörtern, wobei multiple Vererbung nicht stattfindet. Intransitive Verben sind genereller als transitive Verben, weil sie in ihren Beschreibungsstrukturen nur eine Valenz, nämlich die Subjekt-Valenz, aufweisen. Transitive Verben erweitern diese Beschreibung indessen durch eine Objektvalenz (zur Beschreibung von merkmalsbasierten Subsumptionshierar-

$lex.class \in \mathcal{W}$	Wortklassenzugehörigkeit
$lex.lexeme \in \mathcal{L}$	Lexemklassenzugehörigkeit
$lex.valencies \in 2^{\mathcal{V}}$	Valenzrahmen
$lex.features \in 2^{\mathcal{U}}$	morphosyntaktische Eigenschaften
$lex.con \in 2^{\mathcal{F}}$	konzeptuelle Referenzen

Tabelle 3.1: Eigenschaften eines Lexikoneintrags

chien vgl. Carpenter (1992)). Ein Wort wie beispielsweise “*zeigen*” erbt seinerseits die in der Klassenhierarchie definierten Valenzen. Eine Erweiterung des Lexikons erfolgt entsprechend durch die Attribuierung von Wörtern zu Wortklassen, ohne daß für jedes neu eingetragene Wort die Eigenschaften zu spezifizieren wären, da sie geerbt werden.

Für jede im Wortartlexikon spezifizierte Oberflächenform eines Wortes setzen sich die sie vollständig beschreibenden Merkmale aus den im Lexem- und im Wortartlexikon enthaltenen Informationen zusammen. Diese bestehen unter anderem aus den ererbten oder den direkt an einer der entsprechenden Wort- oder Lexemklasse definierten Valenzen, den morphologischen Merkmalen, den Stellungsbeschreibungen sowie den mit dem Lexemeintrag assoziierten Referenzobjekten im Weltwissen, sofern diese vorhanden sind. Das Fehlen eines Identifiers im Weltwissen resultiert einerseits aus der Tatsache, daß Wörter ohne bestimmbar referentiellen Bezug existieren (z.B. Funktionswörter), andererseits kommt auch an dieser Stelle das Problem einer möglichen Unterspezifikation mit ins Spiel.⁶

Zusammenfassend lassen sich die Eigenschaften eines Lexikoneintrags lex als Element der Menge der Lexikoneinträge $\mathcal{LE}$ wie in Tabelle 3.1 beschreiben. Diese Eigenschaften gehen mit der Instantiierung eines Lexikoneintrags auf den ihn repräsentierenden Wortaktor über und dienen als Ausgangspunkt für weitergehende Berechnungen zur Etablierungen von Abhängigkeitsrelationen während des Parsings eines Eingabesatzes. Ausgehend vom aktorenbasierten Parsing (Agha et al., 1997) verwendet SYNDIKATE Wortaktoren für syntaktische Berechnungsprozesse bei der Etablierung von Abhängigkeitsbeziehungen.

⁶Der Zusammenhang von Lexikon und mit lexikalischen Einträgen verknüpften Konzeptidentifiern kommt im nachfolgenden Kap. 3.2.2 ausführlicher zur Darstellung.

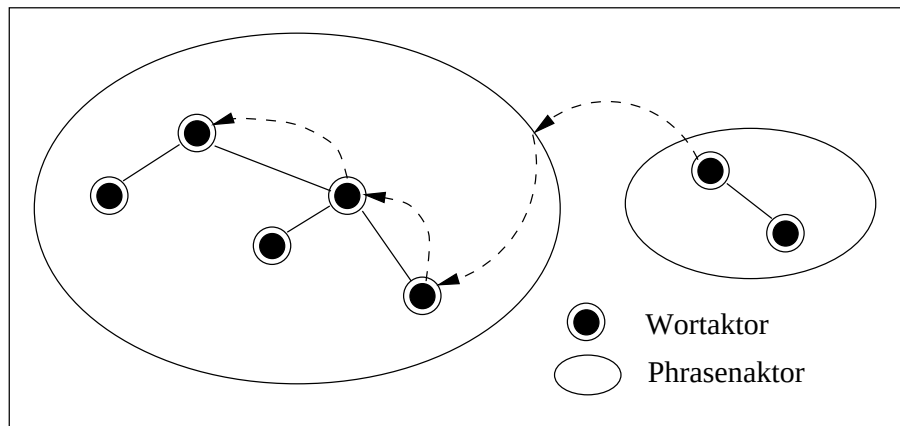


Abbildung 3.4: Interaktion von Wortaktoren und Phrasenaktoren

Versucht nun ein Wortaktor nach der Instantiierung sich syntaktisch an seinen linken Kontext anzubinden⁷, dann sendet er eine Nachricht an andere Wortaktoren, um seinen syntaktischen Kopf zu ermitteln, wobei bei der Adressierung die Erreichbarkeitsbedingungen aus den Wortstellungsdomänen mitberücksichtigt werden. Erhält der anfragende Wortaktor eine positive Nachricht eines anderen Wortaktors, d.h., die Verträglichkeitsbedingungen einer Valenz an einem potentiellen Kopf sind erfüllt, so gehen die beiden beteiligten Wortaktoren eine Dependenzrelation ein. Ergebnisse von Dependenzberechnungen werden von sogenannten *Phrasenaktoren* verwaltet.⁸ Schon bei der Instantiierung eines Wortaktors wird zugleich ein Phrasenaktor mit instantiiert. Darüber hinaus regeln diese auch die Kommunikation zwischen Wortaktoren. Diesen Zusammenhang verdeutlicht Abb. 3.4. In der Darstellung sind zwei Teilberechnungsergebnisse, d.h. zwei partielle Dependenzgraphen, in Phrasenaktoren gekapselt. Beim Versuch eines Wortaktors aus dem rechten Phrasenaktor, seinen syntaktischen Kopf zu bestimmen, geht die Nachricht zunächst an den Phrasenaktor des linken Kontexts. Dieser propagiert die Anfrage innerhalb seines Dependenzgraphen in einer festgelegten, hier indessen nicht relevanten Reihenfolge weiter. Zudem stellen die Phrasenaktoren den Übertragungskanal für Nachrichten zwischen den Wortaktoren dar.

⁷Wegen des inkrementellen Parsings existiert ein rechter Kontext im Satz nicht. Semantisch-konzeptuelle Verträglichkeitsbedingungen seien an dieser Stelle ausgeklammert.

⁸Diese und die nachführenden Ausführungen zu den verschiedenen Akteurenklassen orientieren sich vereinfachend an Neuhaus (1999). Die Darstellung erstreckt sich ausschließlich auf die für diese Arbeit relevanten Sachverhalte.

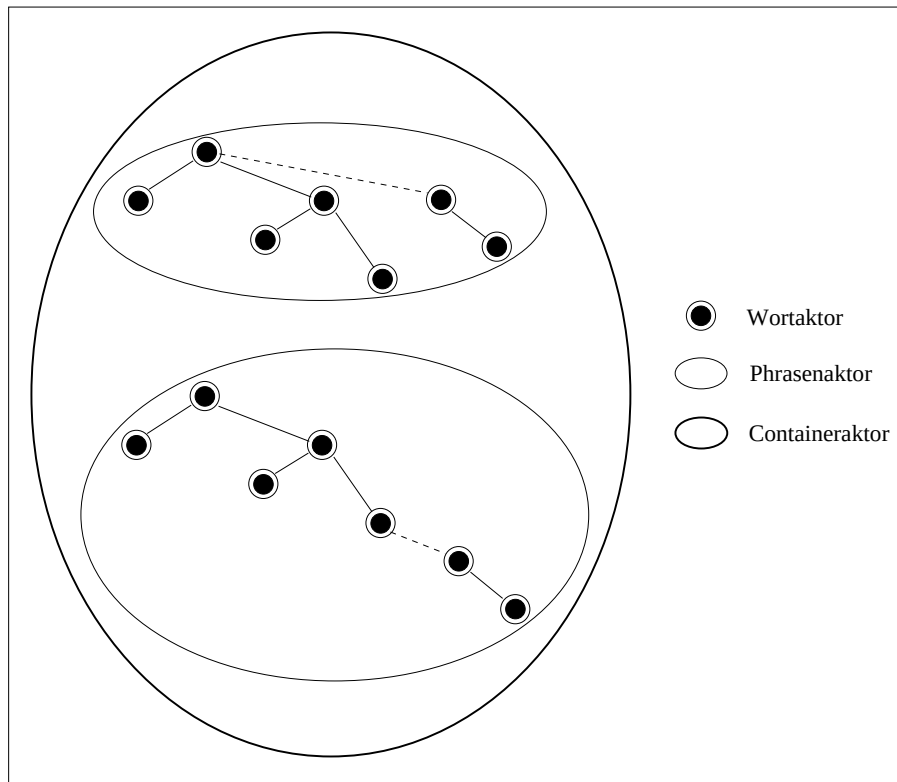


Abbildung 3.5: Duplizierung von Strukturen bei Ambiguitäten und deren Verwaltung durch Containeraktoren

Ein Phrasenaktor kapselt folglich ein Teilberechnungsergebnis während der Analyse eines Eingabesatzes. Darüber hinaus verwaltet ein Phrasenaktor das Kommunikationsverhalten seiner Wortaktoren im Falle von eingehenden Nachrichten von außerhalb liegenden Wortaktoren. Da sich die Wortaktoren, die eine Anfrage miteinander positiv aushandeln, grundsätzlich jeweils innerhalb getrennter Phrasenaktoren befinden, führt die Etablierung der Abhängenzrelation zur Instantiierung eines neuen Phrasenaktors, in dem die Verbindung der beiden Teilabhängenzgraphen über eine Abhängenzrelation zu einem einzigen Abhängenzgraphen vorgenommen wird. Die Vorläufer-Phrasenaktoren bleiben erhalten, um für ein gegebenenfalls notwendiges Backtracking zur Verfügung zu stehen, da der Parser zunächst eine Tiefensuche ausführt⁹.

⁹Aufgrund des restringierten Backtrackings stehen Vorläufer-Phrasenaktoren in der Parsing-Historie nur bis zur Tiefe 2 zur Verfügung.

Es ist jedoch möglich, daß nicht nur ein einziger in einem Phrasenaktor befindlicher Wortaktor eine positive Nachricht auf die Anfrage eines Wortaktors absendet. Liegt dieser Fall vor, dann handelt es sich um eine klassische strukturelle Ambiguität. Da Ambiguitäten im allgemeinen nicht innerhalb eines einzigen Abhängigkeitsgraphen zu repräsentieren sind, wird für jede mögliche Lesart eine Kopie des aktuellen Phrasencontainers erzeugt und jeweils die entsprechende Abhängigkeitsrelation eingezeichnet. Für die adäquate Behandlung von syntaktischen Ambiguitäten sind die Mengen der sie beinhaltenden Phrasenaktoren in einem *Containeraktor* zusammengefaßt. Containeraktoren verwalten die Kommunikation zwischen Phrasenaktoren.

Bei längeren, strukturell reichhaltigeren Sätzen verwendet das Anfrageprotokoll zwischen Wortaktoren folglich die unterschiedlichen Aktorenschichten, um die Kommunikation zu verwalten.

Sei bei einer Nachricht eines Wortaktors die Situation wie in Abb. 3.4 gegeben. Das Verarbeitungsprotokoll leitet dort diese Nachricht an insgesamt drei Wortaktoren weiter. Wenn nun der unterste und der oberste dieser Wortaktoren signalisieren, daß sie eine Abhängigkeitsrelation mit dem anfragenden Aktor eingehen können, reicht die Instantiierung eines einzigen neuen Phrasenaktors zur Kapselung des Analysezustands nicht mehr aus. Stattdessen muß zunächst eine Kopie des linken und rechten Phrasencontainers erzeugt werden, um für beide syntaktische Lesarten eine Strukturbeschreibung bereitzuhalten. Anschließend erfolgt die abhängigkeitsrelationelle Relationierung beider Wortaktoren, für die sämtliche an die Abhängigkeitsrelation geknüpften Bedingungen erfüllt sind, mit dem anfragenden Wortaktor in einem separaten Phrasencontainer (siehe Abb. 3.5 gestrichelte Linien). Die syntaktisch ambigen Lesarten kapselt schließlich der Containeraktor. Kap. 4.3 beschreibt genauer, wie syntaktische mit lexikalischen und semantischen Ambiguitäten interagieren.

Um nun die in der Ausgabe des Parsers enthaltenen syntaktischen Beschreibungsstrukturen nach einer erfolgreichen Satzanalyse zu erläutern, sei folgender Beispielsatz aus der Medizin eingeführt:

1. In einem reiskorngroßen Partikel mit einem Durchmesser von 3 mm wurde eine Magenschleimhaut vom Antrumtyp erfaßt.

Die Eingabe dieses Satzes in das Textverstehenssystem erzeugt den in Abb. 3.6 wiedergegebenen Abhängigkeitsgraphen. Natürlich zeigt die Grafik nur eines der syntaktischen Berechnungsergebnisse, das allerdings eine vollständige Überdeckung des Eingabesatzes

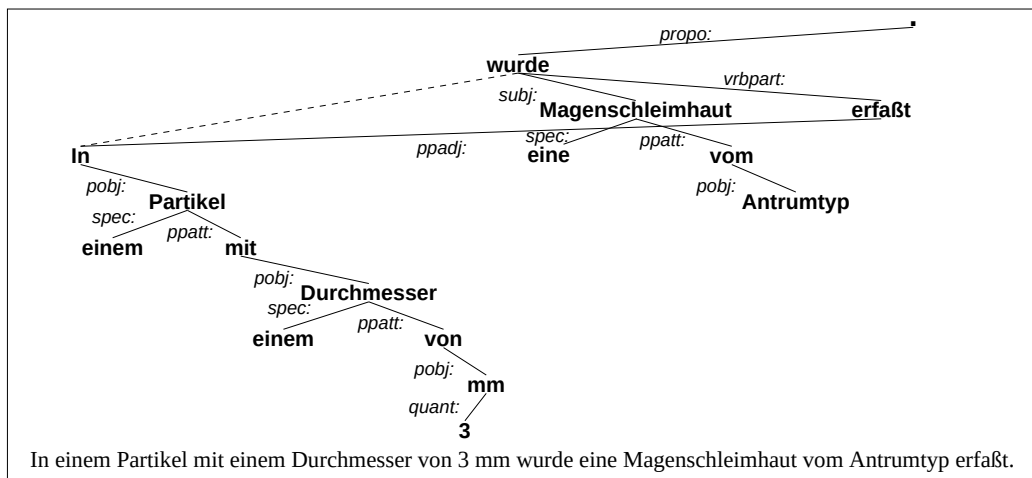


Abbildung 3.6: Dependenzgraph für Satz (1)

charakterisiert. Mit der Konsumierung des Punkts, also der Terminierung der Satzanalyse, finden sich die syntaktischen Lesarten in Phrasenaktoren.

Die Knoten stehen für die Wörter des Eingabesatzes und sind in der Reihenfolge ihres Auftretens von links nach rechts gegeben. Die zwischen den Eingabewörtern bestehenden Dependenzbeziehungen sind über benannte Kanten dargestellt. Der Kopf erscheint in der Darstellung jeweils eine Ebene höher als sein Modifikator, so daß die hierarchischen Abhängigkeiten ebenfalls grafisch gespiegelt sind. So regiert beispielsweise der lexikalische Kopf „*Partikel*“ seinen Modifikator „*mit*“ über die Dependenzrelation *ppatt* (Abk. für Präpositionalattribut). Die im Dependenzgraphen enthaltene gestrichelte Linie verdeutlicht, wie Wortstellungsdomänen den Parsingprozeß steuern. Zwar ist „*erfaßt*“ der syntaktische Kopf der Präpositionalphrase „*in einem Partikel*“, die Positionszuweisung der Präpositionalphrase im Dependenzgraphen erfolgt indessen durch den syntaktischen Kopf von „*erfaßt*“, nämlich „*wurde*“. Schließlich wurde das im Eingabesatz vorhandene Wort „*reiskorn groß*“ nicht in die Analyse integriert, da es im Lexikon nicht enthalten ist.

Jedoch zielt SYNDIKATE im Kern nicht darauf ab, nur syntaktische Strukturen zu erzeugen, vielmehr wird eine inhaltliche Interpretation der Dependenzgraphen zugleich mit der syntaktischen Analyse vorgenommen. Das hierzu notwendige Wissen wird im folgenden Unterkapitel besprochen.

3.2.2 Wissensrepräsentationssprache, Weltwissen und Semantik

Sowohl das Textverstehen in SYNDIKATE als auch die dem Textverstehen zugrundeliegenden Inferenz-Prozesse sind stark wissensbasiert. Das heißt, das Textverstehen hängt qualitativ unmittelbar von den nachfolgenden Merkmalen des zur Verfügung stehenden Wissens ab:

- von der Abdeckung der betrachteten Teildomäne (*coverage*) durch eine ausreichende Menge an Termen und Konzepten; das Coverage-Problem stellt wegen der enormen Anzahl an Konzepten in einigen Domänen wie zum Beispiel der Medizin oder der Pharmakologie eine begrenzende Größe für jedes wissensbasierte Textverstehenssystem dar, das für derartige Domänen konzipiert wird;
- von der Tiefe des Domänenmodells (*granularity*), d.h. der Reichhaltigkeit an Relationierungen zwischen Termen bzw. Konzepten und der Differenzierung eher genereller Konzeptbeschreibungen wie etwa COMPUTER-SYSTEM, welche durch eine größere Anzahl an detaillierter definierten Spezialisierungen (z.B. NOTEBOOK, WORKSTATION, PC etc.) erweitert werden können, sowie schließlich
- von der Güte des bereitgestellten Wissens, d.h. der Konsistenz des Domänenmodells bezüglich formaler Anforderungen an das Begriffssystem, wie etwa die Zyklenfreiheit von Subsumptionsbeziehungen (Beschreibungsebene) oder die Ableitbarkeit semantischer Beschränkungen für die Relationierung von Termen (Reasoningebene).

Die Fragen, welche Arten von Wissen zum Textverstehen bzw. Reasoning über Textpropositionen benötigt werden bzw. ob und inwieweit diese Wissensarten darstellbar sind sowie wie sich die konkrete Implementation in SYNDIKATE gestaltet, wird ausführlich in Kapitel 5 behandelt. An dieser Stelle sollen nur die für den Systemüberblick und für das zitierte Beispiel notwendigen formalen Beschreibungen eingeführt werden.

Der in SYNDIKATE verwendeten Wissensrepräsentationssprache liegt eine terminologische Logik zugrunde. Terminologische Logiken entstanden aus der Kritik am epistemologischen Status von Links in den früher überaus populären semantischen Netzen (vgl. Brachman (1979), Brachman (1983)). Im Gegensatz zu den semantischen Netzen weisen terminologische Logiken eine formale Fundierung in der Modelltheorie auf. D.h., es kann für Konzeptbeschreibungen ermittelt werden, ob sie bezüglich der terminologischen Axiome einer Wissensbasis erfüllbar sind (Baader et al., 1992).

Syntax	Semantik	Syntax	Semantik
$\top$	$\Delta^{\mathcal{I}}$	R_{atom}	$\{(d, e) \in \Delta \times \Delta \mid (d, e) \in R^{\mathcal{I}}\}$
C_{atom}	$\{d \in \Delta \mid d \in C^{\mathcal{I}}\}$	$R \sqcap Q$	$R^{\mathcal{I}} \cap Q^{\mathcal{I}}$
$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$	$R \sqcup Q$	$R^{\mathcal{I}} \cup Q^{\mathcal{I}}$
$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$	$c R$	$\{(d, d') \in R^{\mathcal{I}} \mid d \in C^{\mathcal{I}}\}$
$\neg C$	$\Delta \setminus C^{\mathcal{I}}$	$R _C$	$\{(d, d') \in R^{\mathcal{I}} \mid d' \in C^{\mathcal{I}}\}$
$\forall R.C$	$\{d \in \Delta \mid R^{\mathcal{I}}(d) \subseteq C^{\mathcal{I}}\}$	R^{-1}	$\{(d, d') \in \Delta \times \Delta \mid (d', d) \in R^{\mathcal{I}}\}$
$\exists R.C$	$\{d \in \Delta \mid R^{\mathcal{I}}(d) \cap C^{\mathcal{I}} \neq \emptyset\}$	$(R_1, \dots, R_n)$	$R_1^{\mathcal{I}} \circ \dots \circ R_n^{\mathcal{I}}$
$\exists_{\geq n} R$	$\{d \in \Delta \mid R^{\mathcal{I}}(d) \geq n\}$		
$\exists_{\leq n} R$	$\{d \in \Delta \mid R^{\mathcal{I}}(d) \leq n\}$		
$A_1 \parallel \dots \parallel A_n$	$\llbracket A_i \rrbracket^{\mathcal{I}} \cap \llbracket A_j \rrbracket^{\mathcal{I}} = \emptyset, i \neq j$		

Tabelle 3.2: Syntax und Semantik von terminologischen Konzept- und Rollentermen

Zur formalen modelltheoretischen Beschreibung der Terminologischen Logik wird die Interpretation $\mathcal{I}$ verwendet, mit $\mathcal{I} = \langle \Delta, \llbracket \cdot \rrbracket^{\mathcal{I}} \rangle$, wobei Δ eine beliebige Menge (Domäne) bezeichnet und $\llbracket \cdot \rrbracket^{\mathcal{I}}$ eine Interpretationsfunktion. Konzepte sind unäre Prädikate über eine Domäne Δ , während Relationen binäre Prädikate über eine Domäne Δ darstellen. Schließlich konstituieren Symbole für Individuen die Elemente in Δ . Die Interpretationsfunktion $\llbracket \cdot \rrbracket^{\mathcal{I}}$ weist nun jedem Konzeptsymbol $C \in \mathcal{F}$ eine Teilmenge der Domäne Δ zu: $\llbracket \cdot \rrbracket^{\mathcal{I}} : \mathcal{F} \rightarrow 2^{\Delta}$; wobei $\mathcal{F}$ die Menge aller Konzeptsymbole bezeichnet. Entsprechend wird jedem Rollensymbol $R \in \mathcal{R}$ eine Menge bestehend aus Tupeln von Elementen in Δ zugewiesen: $\llbracket \cdot \rrbracket^{\mathcal{I}} : \mathcal{R} \rightarrow 2^{\Delta \times \Delta}$; wobei $\mathcal{R}$ die Menge aller Relationssymbole bezeichnet. Zu guter Letzt sind die Individuen $I \in \mathcal{I}$ die Elemente von Δ unter der Abbildung durch die Interpretationsfunktion $\llbracket \cdot \rrbracket^{\mathcal{I}} : \mathcal{I} \rightarrow \Delta$, wobei $\mathcal{I}$ genau die Menge der Individuensymbole bezeichnet.

Die Tabelle 3.2 enthält ein Fragment der terminologischen Sprache $\mathcal{CDL}$ (Heinsohn et al., 1994), deren modelltheoretische Semantik unter der Interpretation $\mathcal{I}$ aufgezeigt wird. Die linke Seite der Tabelle beinhaltet Ausdrücke zur Definition von Konzepten, während sich die rechte Seite aus mengentheoretischen Ausdrücken zur Definition von Relationen zusammensetzt. Dabei sind die Symbole C und D Konzeptbezeichner ($C, D \in \mathcal{F}$) sowie Q und R Relationsbezeichner ($R, Q \in \mathcal{R}$). Jeder der beiden Tabellenteile faßt die Syntax und Semantik von konzept- bzw. relationsformenden Ausdrücken einer Terminologischen

Logik zusammen. Bei der Definition der Semantik kommt für Relationen eine funktionale Schreibweise der Form $R^{\mathcal{I}}(d) = \{e \mid (d, e) \in R^{\mathcal{I}}\}$ zur Anwendung. Aus den atomaren Beschreibungen für Konzepte (links) bzw. Relationen (rechts) lassen sich induktiv über Operatoren neue Konzeptbeschreibungen gewinnen. Für Konzepte sind dies (in der Reihenfolge des Auftretens in Tabelle 3.2) die Konjunktion, Disjunktion, die Negation, die allquantifizierte Typbeschränkung für eine Rolle R , die existenzquantifizierte Typbeschränkung für eine Rolle R , numerische Beschränkungen für die Anzahl von Füllern einer Rolle R nach oben und nach unten. Als “syntactic sugar” kommt noch die Definition einer sogenannten Partition hinzu, bei der alle $n * (n - 1)$ Konzeptpaare unter der Interpretation $\mathcal{I}$ jeweils kein gemeinsames Element haben, wobei sich dies auch durch intensiven Gebrauch der Negation darstellen ließe. Bei den Relationen gibt es neben der atomaren Form ebenfalls die Konjunktion und die Disjunktion. Darüber hinaus lassen sich Typenrestriktionen für den Domänen- und den Wertebereich festlegen, die Inverse zu einer Relation definieren sowie Relationen über Komposition aus Relationen bilden.

Das dargestellte induktive Verfahren dient als Grundlage zur Formulierung terminologischer Axiome. Terminologische Axiome stellen entweder Konzeptdefinitionen durch Zuweisung eines Symbols zu einem terminologischen Ausdruck dar oder aber Konzeptspezialisierungen durch Teilmengenbeziehungen. Bei der Konzeptdefinition müssen notwendige und hinreichende Bedingungen gegeben werden können, wohingegen bei der Konzeptspezialisierung notwendige Bedingungen genügen. In $A \doteq C$ weist der Operator $\doteq$ einem Konzeptbezeichner A eine Konzeptbeschreibung in Form eines konzeptformenden Ausdrucks C zu, wenn hinreichende und notwendige Bedingungen zur Konzeptdefinition vorliegen, wobei C entweder atomar ist oder aus einem komplexeren konzeptformenden Ausdruck besteht (siehe Tabelle 3.3). Durch Verwendung des Operators $\sqsubseteq$ in dem Ausdruck $A \sqsubseteq C$ hingegen wird angezeigt, daß es nur notwendige Bedingungen gibt, um A durch C zu definieren. Die Definition von terminologischen Axiomen über Relationen erfolgt auf analoge Weise mit denselben Operatoren (ebenfalls Tabelle 3.3).

Unglücklicherweise wird der $\sqsubseteq$ Operator oftmals ebenfalls dazu verwendet, um eine inferierte (d.h. aus den formal-logischen Beschreibungen ableitbare) Subsumptionsbeziehung anzuzeigen. Zur Vermeidung dieses Ambiguitätsproblems beim Gebrauch des $\sqsubseteq$ -Operators verwende ich zum Kennzeichen einer inferierten Subsumptionsbeziehung den $\preceq$ -Operator. Eine endliche Menge an terminologischen Axiomen nennt man eine Terminologie $\mathcal{T}$ oder T-Box. Ferner gilt, daß ein Konzeptbezeichner nur einmal auf der linken Seite eines ter-

Terminologische Axiome	Semantik	Assertionale Axiome	Semantik
$A \doteq C$	$A^I = C^I$	$a : C$	$a^I \in C^I$
$A \sqsubseteq C$	$A^I \subseteq C^I$	$a R b$	$(a^I, b^I) \in R^I$
$Q \doteq S$	$Q^I = S^I$		
$Q \sqsubseteq R$	$Q^I \subseteq R^I$		

Tabelle 3.3: Syntax und Semantik von terminologischen und assertionalen Axiomen

ANIMAL	$\sqsubseteq$	ANIMATE	(1)
VERTEBRATE	$\doteq$	ANIMAL $\sqcap \exists$ HAS-ANATOMICAL-PART.SPINAL-COLUMN	(2)
FISH	$\doteq$	ANIMAL $\sqcap \exists$ HAS-ANATOMICAL-PART.SPINAL-COLUMN $\sqcap$ $\forall$ HAS-HABITAT.WATER	(3)
FISH	$\preceq$	VERTEBRATE	(4)

Tabelle 3.4: Beispiel einer Menge an terminologischen Axiomen

minologischen Axioms erscheinen darf, um die Eindeutigkeit für Konzeptbezeichner zu gewährleisten.

Tabelle 3.4 enthält ein kleines Fragment einer T-Box zur Veranschaulichung der Verwendung der Operatoren $\sqsubseteq$, $\doteq$ und $\preceq$. Nach (1) definiert sich das Konzept ANIMAL notwendig als Subkonzept von ANIMATE ($\sqsubseteq$ -Operator). Dem Konzept, das dadurch beschrieben ist, zum Konzept ANIMAL zu gehören und mindestens einen Füller der Relation HAS-ANATOMICAL-PART vom Typ SPINAL-COLUMN zu haben (2), wird der Konzeptidentifizier VERTEBRATE zugewiesen ($\doteq$ -Operator). Dies sei der terminologische Ausdruck dafür, daß sich ein Wirbeltier hinreichend und notwendig durch das Vorhandensein eines anatomischen Teils vom Typ Wirbelsäule beschreiben läßt. Zur Definition des Konzepts FISH tritt zu diesen Axiomen, daß das Habitat der Fische durch Verwendung des Allquantors ausschließlich das Wasser ist (3). Um differenzieren zu können, daß sich die Subsumptionsbeziehung zwischen VERTEBRATE und FISH nicht per definitionem durch einen intentionalen Akt des Knowledge Engineers ergibt wie in (1), sondern aus den Konzeptbeschreibungen selbst inferierbar ist, kommt in (4) der $\preceq$ -Operator zur Anwendung.

Mit dem bisher dargestellten T-Box-Mechanismus können jedoch nur Aussagen über Mengen gemacht werden, nicht aber über einzelne Elemente dieser Mengen. Deshalb existiert neben der rein terminologischen Ebene noch eine assertionale Ebene, in der Aussagen über Individuen oder auch *Instanzen* innerhalb eines formal definierten Logikformalismus gemacht werden. Tabelle 3.3 beschreibt auf der rechten Seite die Syntax und die formale Semantik der assertionalen Axiome unter der Interpretation $\mathcal{I}$. Wird ein Symbol a als Instanz einem Konzeptsymbol C zugewiesen, dann ist die Interpretation dieses Symbols genau ein Element in der Menge der Interpretation des Konzeptsymbols. Analog hierzu sind zwei Instanzsymbole a und b genau dann über ein Relationssymbol miteinander relationiert, wenn die Interpretation des Symboltupels in der Interpretation des Relationssymbols liegt. Die Eindeutigkeit eines Instanzsymbols wird durch die sogenannte *unique name assumption* garantiert: seien a und b Instanzsymbole, dann gilt $a^{\mathcal{I}} \neq b^{\mathcal{I}}$.

Die hiermit eingeführte assertionale Ebene bezeichnet man in Absetzung zur T-Box auch als A-Box, die eine endliche Menge $\mathcal{A}$ solcher Objekt- und Rollenbeschreibungen beinhaltet. Einen Formalismus, der sowohl terminologische als auch assertionale Axiome aufweist, bezeichnet man auch als Hybridformalismus. Eine Interpretation $\mathcal{I}$ ist ein Modell für eine A-Box bezüglich einer gegebenen T-Box genau dann, wenn $\mathcal{I}$ die assertionalen und terminologischen Axiome erfüllt.

Terminologische Logiken definieren zwar den formalen Rahmen, in dem Konzepte und Relationen beschrieben werden können. Aber für die konkrete Implementation von Terminologien im Rahmen einer computergestützten Anwendung wie dem Textverstehenssystem SYNDIKATE rücken die praktischen Aspekte der Anwendung derartiger Beschreibungssysteme in den Vordergrund. Terminologische Logiken als Wissensrepräsentationssprachen stehen auf der Anwendungsseite in Form von sogenannten KL-ONE-artigen Wissensrepräsentationssystemen zur Verfügung, von denen eine ganze Reihe mit jeweils unterschiedlicher Ausdrucksmächtigkeit existiert (Woods & Schmolze, 1992). In SYNDIKATE kommt das hybride Wissenrepräsentationssystem LOOM zum Einsatz (MacGregor, 1994), das auf der terminologischen Ebene eine relativ reiche Sprache zur Konzeptdefinition aufweist (Horrocks, 1995) sowie auf der assertionalen Ebene in etwa die Sprache $\mathcal{AF}$ abdeckt (Nebel, 1990a).

Sowohl in der T-Box als auch in der A-Box sind besondere Schlußfolgerungsmechanismen vorhanden, die aus den terminologischen Axiomen Subsumptionsbeziehungen ableiten und aus den Assertionen die Zugehörigkeit von Instanzen zu Konzepten berechnen. Mit

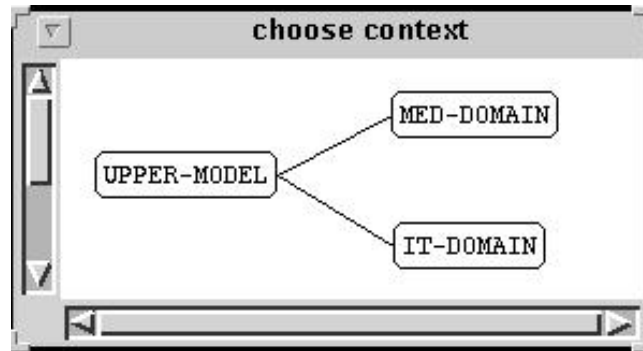


Abbildung 3.7: Vererbungshierarchie zwischen Wissensbasen

Hilfe von Konzeptdefinitionen durch vom Wissenrepräsentationssystem bereitgestellte Ausdrücke und der modelltheoretischen Semantik lassen sich diese Subsumptionsbeziehungen automatisch von einer deduktiven Reasoning-Komponente, dem sogenannten Classifier, berechnen (MacGregor & Brill, 1992). Weiterhin ist der Classifier in der Lage, inkonsistente Konzeptdefinitionen zu erkennen und gewährleistet damit ein kohärentes Modell des Weltausschnitts. Bedingt durch die Komplexitätseigenschaften bei der Berechnung von Subsumptionsbeziehungen sind die Classifier i.a. nicht vollständig (Nebel, 1990b). Es gibt hierbei einen Trade-Off zwischen Ausdrucksmächtigkeit und Vollständigkeit der Klassifikationsalgorithmen (Baader, 1991). Eine vertiefende Auseinandersetzung mit Wissensrepräsentationssprachen in Zusammenhang mit der Darstellbarkeit von Wissen findet sich in Abschnitt 5.1, weshalb der Problembereich an dieser Stelle nicht weiter vertieft wird.

Das in SYNDIKATE vordefinierte Wissen existiert einerseits in Form einer domänenunabhängigen Teilwissensbasis $\mathcal{KB}_{upper}$, das sogenannte UPPER-MODEL, welches allgemeingültige und abstrakte Konzepte und Relationen mittels terminologischer und assertionaler Axiome definiert. Es gilt daher $\mathcal{KB}_{upper} := \mathcal{T}_{upper} \cup \mathcal{A}_{upper}$. Die jeweiligen an die Spezialdomänen Informationstechnologie und Medizin angepassten Wissensbasen $\mathcal{KB}_{it}$ bzw. $\mathcal{KB}_{med}$ fügen diesem vordefinierten Wissen weitere domänenspezifische terminologische und assertionale Axiome hinzu. Damit ergibt sich für die Informationstechnologiedomäne $\mathcal{KB}_{it} := \mathcal{KB}_{upper} \cup \mathcal{T}_{it} \cup \mathcal{A}_{it}$ und für die Medizindomäne $\mathcal{KB}_{med} := \mathcal{KB}_{upper} \cup \mathcal{T}_{med} \cup \mathcal{A}_{med}$. Abb. 3.7 verdeutlicht diesen Zusammenhang nochmals grafisch, indem zwischen den Wissensbasen eine Vererbungshierarchie dargestellt wird. Die Upper-Model-Wissensbasis ist folglich sowohl Bestandteil der IT-Domain-Wissensbasis als auch der Med-Domain-Wissensbasis.

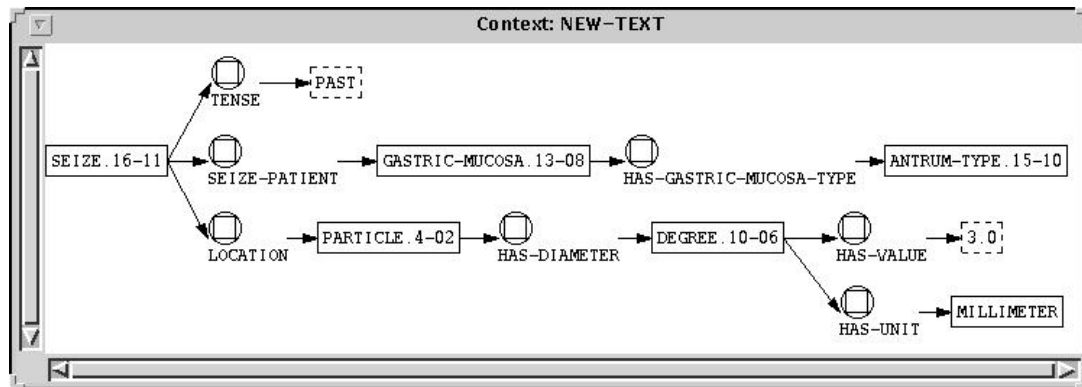


Abbildung 3.8: Repräsentation von Satz (1) durch Konzeptgraphen

Die Funktion des in LOOM implementierten allgemeinen Wissens und des Spezialwissens für die Medizin und die Informationstechnologie ist eine doppelte. Erstens dient eine exakte Kopie der Domänenwissensbasis der inkrementell aufgebauten Textwissensbasis als Interpretationskontext, indem das durch die Domänenwissensbasis gegebene Modell sukzessive um assertionale Axiome erweitert wird. Zweitens sind aus dem konzeptuellen Wissen Typrestriktionen für die Relationierbarkeit von Diskursobjekten in der Textwissensbasis ableitbar. Die Projektion von syntaktischen Strukturen in eine Textwissensbasis erfolgt zunächst über die Instantiierung der mit Wortoberflächen verbundenen Referenzobjekte im Weltwissen. Dazu wird entsprechend der im Lexikon vorhandenen Information über die Zugehörigkeit einer Wortoberfläche zu einem Konzept C oder mehreren Konzepten eine Assertion des Typs $(a : C)$ der Textwissensbasis hinzugefügt. Damit wird dem Problem der Referenz in Texten insofern Rechnung getragen, als der Einführung eines neuen Diskursobjekts im Text eine dazugehörige Instanz in der Textwissensbasis zugeordnet wird, ohne weitergehende Aussagen darüber zu machen, ob es tatsächlich um ein Individuum oder um generische Aussagen über Mengen an Objekten geht (vgl. hierzu Azzam (1994)). Im Fortgang der Analyse werden unter Berücksichtigung syntaktischer Wohlgeformtheitsbedingungen durch die semantische Interpretation die konzeptuellen Relationierungen zwischen diesen Instanzen berechnet und ebenfalls als Assertionen des Typs $(a R b)$ in die Textwissensbasis eingestellt, so daß aus einem Dependenzgraphen durch die semantische Interpretationsschicht eine inhaltliche Repräsentation in Form eines Konzeptgraphen erzeugt wird.

Für Satz (1) führt dieser Ansatz zur Instantiierung aller im Lexikon mit einem Konzept-identifizier verbundenen Wortoberflächen sowie der Relationierung dieser Instanzen über konzeptuelle Relationen. Abb. 3.8 enthält die inhaltliche Repräsentation des Eingabesatzes (1) in Form eines Konzeptgraphen in dem für diese allgemein üblichen Darstellungsformat (vgl. z.B. Sondheimer et al. (1984)), das assertionale und terminologische Axiome graphisch repräsentiert. Instanzen sind durch rechteckige Boxen gegeben sowie durch einen eindeutigen String (“unique name”) qualifiziert, der sich aus dem zur Wortoberfläche gehörenden Konzeptbezeichner zur Instantiierungszeit und einem systemgenerierten, durch einen Punkt angeschlossenen numerischen String zusammensetzt (vgl. Abb. 3.8 GASTRIC-MUCOSA.13-08). Relationen zwischen Instanzen bestehen aus einer gerichteten und gelabelten Kante zwischen diesen Instanzen (siehe Abb. 3.8 GASTRIC-MUCOSA.13-08 HAS-GASTRIC-MUCOSA-TYPE ANTRUM-TYPE.15-10). Darüber hinaus werden noch Symbole verwendet, um Zeitmarken oder Modalitätsmarken mit Instanzen zu relationieren. Symbole unterscheiden sich grafisch von Instanzen dadurch, daß das sie umgebende Rechteck gestrichelt ist (SEIZE.16-11 TENSE 'PAST). Eine über das Zuweisen dieser Marken hinausgehende tiefere Interpretation von Modalität und Tempus findet beim augenblicklichen Stand der Implementation nicht statt.

Der abgebildete Konzeptgraph beinhaltet eine vereinfachte inhaltliche Darstellung des “komplexeren” Dependenzgraphen. Die Abbildung von der Ebene der syntaktischen Repräsentation auf die Ebene der Konzeptgraphen wird hierbei inkrementell durch die Regeln zur Semantischen Interpretation geleistet (vgl. Kap. 4.1).

Darüber hinaus weist das von uns verwendete Wissensrepräsentationssystem LOOM noch eine Besonderheit auf (MacGregor & Yen, 1989), die für das Schlußfolgern über aus den syntaktischen Strukturen gewonnenen konzeptuellen Repräsentationen unerlässlich ist: sein Produktionsregelsystem (Yen et al., 1991). Dieses registriert jede in der Textwissensbasis hinzugefügte Assertion. Bei Erfüllung von in den Regeln im Bedingungsteil kodierten Assertionsmustern können mittels eines Aktionsteils Veränderungen in der A-Box herbeigeführt (z.B. die Löschung von Assertionen) oder weitere Prüfungen von Constraints ausgeführt werden. Letzteres spielt bei der in Kapitel 4.1.3 besprochenen *konzeptuellen Interpretation* eine wichtige Rolle.

Ein besonderes Merkmal des Ansatzes zur semantischen Interpretation in SYNDIKATE besteht darin, daß die semantische Interpretation in demselben Formalismus behandelt wird wie die Repräsentation des Weltwissens (terminologische Logik). Der Verzicht

auf eine intermediäre Repräsentationsschicht, die ausschließlich der semantischen Interpretation gewidmet ist, führt im Resultat zu Einschränkungen der Ausdrucksmächtigkeit. In der Tat sind nicht alle sprachlichen Phänomene, die inhaltsrelevant sind, durch diese Einschränkungen repräsentierbar. Gegenüber formal ausgerichteten Ansätzen zur semantischen Interpretation ergeben sich hier deutliche Defizite. Jedoch beschäftigt sich diese Arbeit schwerpunktmäßig damit, wie heterogene Wissensquellen in einem realen Anwendungsszenario zur Wissensextraktion zusammenarbeiten, um ein Maximum von Fakten aus Texten zu extrahieren. Um jedoch die Einschränkungen zu dokumentieren und Begründungen für diese zu liefern, beschäftigt sich Kap. 5.1 ausführlich mit diesem Thema.

3.2.3 Versionierungslerner und Konzeptlerner

Das für das Textverstehenssystem SYNDIKATE modellierte Domänenwissen unterliegt in unterschiedlichem Maße Veränderungen. Während das domänenunabhängige Upper-Model und Bereiche der Medizin, wie etwa die Anatomie, in dem Sinne statisch sind, daß elementare Erweiterungen oder weitreichende Veränderungen nicht zu erwarten sind, zeichnet sich die Entwicklung der Informationstechnologie durch eine ungebrochene Dynamik aus. Ebenso existieren in der Medizindomäne forschungsintensive Bereiche. Dabei sind grundsätzlich zwei Fälle zu unterscheiden.

Zunächst kann sich die Definition eines bereits spezifizierten Konzepts wandeln, ein Phänomen, das als *Concept Drift* bezeichnet wird (Klenner & Hahn, 1994). Beispielhaft läßt sich dies am Konzept COMPUTER-SYSTEM belegen. Kamen die ersten Heimcomputer noch ohne Festplatte, werden zu Ende der neunziger Jahre alle Systeme dieser Art mit Festplatten ausgeliefert, von deren Speicherkapazität selbst Administratoren großer Rechnetze in den Achtzigern nur träumen konnten. Allerdings gibt es heute auch wieder eine Gegenbewegung, die Clients ohne Festplatten konfiguriert, um den Administrationsaufwand zu minimieren. Um das Konzeptsystem nicht ständig von Hand aktualisieren zu müssen, steht SYNDIKATE ein Versionierungslerner zur Verfügung (Hahn & Klenner, 1994). Ausgangspunkt des Versionierungslerners sind die innerhalb des Domänenwissens vorgehaltenen Konzeptdefinitionen. Diese legen Wertebereiche für einzelne Rollen eines Konzepts fest, wie beispielsweise eine numerische Obergrenze für die RAM-Bestückung eines Computersystems. Beinhalten nun die analysierten Texte Wertebereichsverletzungen, so sind diese ein Indikator dafür, daß die entsprechende Konzeptdefinition im Weltwissen

aktualisiert werden muß. Über das fortlaufende Protokollieren von Constraint-Verletzungen schließlich werden Konzeptdefinitionen stetig auf einen aktuellen Stand gebracht.

Darüber hinaus führen sowohl grundlegende Neuerungen, wie etwa das Entdecken eines neuen Krankheitserregers oder die Entwicklung eines neuen Produkts, als auch die systeminhärente Unterspezifikation dazu, daß kontinuierlich Lücken im Weltwissen gefüllt werden müssen. Dazu wird in SYNDIKATE das Domänenwissen automatisch durch einen Konzeptlerner erweitert (Hahn & Schnattinger, 1998). Mithilfe des sprachlichen Kontexts und des vorhandenen Domänenwissens, in dem ein zu lernendes Konzept verwendet wird, baut die Lernkomponente AQUA Hypothesen über dessen Eigenschaften auf und bewertet diese durch einen auf Terminologischer Logik basierenden Qualitätskalkül (Hahn et al., 1996a; Schnattinger, 1998). Sind ausreichende Evidenzen für die Präferenz einer Lernhypothese innerhalb des Textmaterials aufgetreten, wird die entsprechende Lernhypothese selektiert, alle anderen hingegen werden verworfen. Das dabei akquirierte Wissen wird anschließend in das bestehende Domänenwissen integriert und steht damit nachfolgenden Textanalysen zur Verfügung.

3.2.4 Skizze des Parsingprozesses

In diesem Unterkapitel wird kurz skizziert, wie SYNDIKATE für einen Eingabesatz eine Abhängigkeitsrelation etabliert und wie die einzelnen Wissenskomponenten dabei interagieren, d.h., wie sich der Kontrollfluß zwischen den heterogenen Wissenskomponenten gestaltet. Da der Lerner in den Standardfall nicht involviert ist, sondern eine Erweiterung der Architektur darstellt, wird er hier nicht erörtert.

Der Parser verarbeitet die Eingabe –wie oben bereits angeführt– in Analogie zum menschlichen Sprachverstehen inkrementell von links nach rechts. Die Schnittstelle zwischen dem Eingabetext und SYNDIKATE ist der Wort-Scanner (siehe Abb. 3.9, A_1), der die im Text T_i enthaltenen Wörter sukzessive dem Analysemechanismus zuführt. Nach dem Einlesen eines Wortes prüft der Scanner, ob das eingelesene Wort im Lexikon enthalten ist (siehe Abb. 3.9, A_2).¹⁰ Im positiven Fall erfolgt die Instantiierung eines entsprechenden

¹⁰Da die hier besprochenen Analyseschritte domänenunabhängig sind, wird hier und im folgenden allgemein vom Lexikon oder der Wissensbasis gesprochen. Dennoch verdeutlicht die Abb. 3.9 die in Kap. 3 beschriebene Philosophie eines domänenunabhängigen Systemkerns mit den entsprechenden domänenspezifischen Spezialisierungen für die Wissensmodule Lexikon und Weltwissen (als Dreiecke dargestellt).

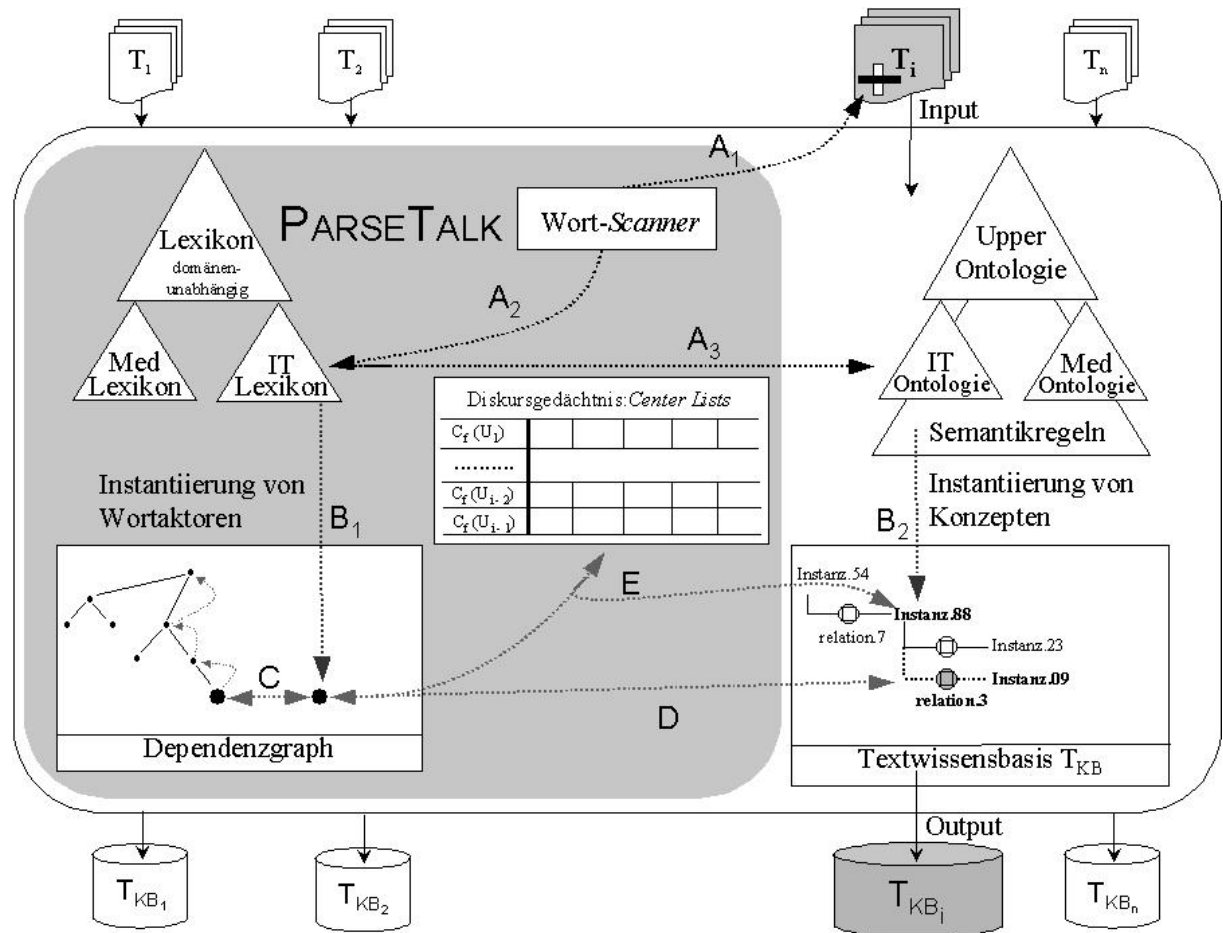


Abbildung 3.9: Interaktion der Wissensquellen bei der Textanalyse (Kontrollfluß)

Wortaktors auf der Seite des Parsers (Abb. 3.9, B_1). Ist darüber hinaus mit dem lexikalischen Eintrag ein Konzeptidentifer assoziiert (Abb. 3.9, A_3), so wird auch vom Konzepttyp dieses Identifiers eine Instanz erzeugt und in die Textwissensbasis eingestellt (Abb. 3.9, B_2).

Der neu instantiierte Wortaktor versucht nun, sich an den bisher aufgebauten Dependenzgraphen anzubinden, indem er am direkten linken Kontext anfragt, ob die Etablierung einer Dependenzbeziehung zulässig ist (Abb. 3.9, C). Dabei läuft die Anfragerichtung immer vom Modifikator zum lexikalischen Kopf. Darüber hinaus wird die Anfrage des Modifikators vom angefragten Kopf im bereits bestehenden Dependenzgraphen entlang der rechten Kante nach oben propagiert, was durch die kleinen Pfeile zwischen den aufsteigenden Knoten im Dependenzgraphen versinnbildlicht wird.

Um eine Dependenzbeziehung eingehen zu können, müssen eine Reihe von Constraints

erfüllt sein: Der tentative syntaktische Kopf muß über eine Valenz für den anfragenden Modifikator verfügen. Ebenso müssen die Kongruenzbedingungen bezüglich der Merkmalsstrukturen erfüllt sein sowie definierte Stellungsrestriktionen über die Position von Wörtern im aktuellen syntaktischen Kontext. Nur wenn alle diese Bedingungen erfüllt sind, wird als letzter Schritt eine Überprüfung auf konzeptuelle Verträglichkeit auf der Weltwissenseite durchgeführt (Abb. 3.9, *D*). Dieser letzte Schritt der Constraint-Prüfung beinhaltet mehr als einen einfachen Type-Check, denn die aus dem syntaktischen Kontext ableitbaren Beschränkungen bezüglich der konzeptuellen Relationierbarkeit der involvierten Instanzen werden im Weltwissen mit berücksichtigt (vgl. Kapitel 4). Zugleich ist der zulässige Typ einer Relation nicht auf deren unmittelbaren Füller restringiert, sondern wird aufgrund von wohldefinierten Erreichbarkeitskriterien innerhalb des Konzeptgraphen selektiert. Kann für eine angefragte Relationierung keine gültige semantische Interpretation ermittelt werden, unterbleibt die Relationierung auf syntaktischer wie konzeptueller Seite. Durch diesen semantischen Filter entfällt der Aufbau syntaktisch valider, konzeptuell jedoch nicht zulässiger Dependenzgraphen. Im positiven Falle werden die entsprechenden Relationierungen auf beiden Seiten eingezogen.

Da die Transition von syntaktischen zu konzeptuellen Relationen und die dabei ablaufenden Reasoningprozesse im Zentrum dieser Arbeit stehen, wird für die eigentliche semantische Interpretation auf Kapitel 4 verwiesen.

Daß bei der syntaktischen Anbindung eines Wortaktors innerhalb eines Dependenzgraphen möglicherweise textuelle Phänomene berücksichtigt werden müssen, wird im unmittelbar nachfolgenden Kapitel 3.2.5 ausführlich diskutiert. An dieser Stelle sei nur bereits angemerkt, daß auf der Systemebene die Protokolle zur Auflösung von referentiellen Bezügen i.a. ebenfalls schon bei den Berechnungen von Constraints für eine Dependenzrelation mitberücksichtigt werden. Dazu wird der entsprechende Referent aus dem Diskursgedächtnis für den vorherigen Satz abgerufen und in die laufenden Reasoningprozesse mit einbezogen (Abb. 3.9, *E*) - vorausgesetzt, die Triggerbedingungen für die Anaphernaufflösung sind erfüllt. Die Beschreibung, wie Textphänomene resoliert werden und das Diskursgedächtnis aufgebaut ist, werden im nachfolgenden Kap. 3.2.5 ausführlich besprochen und deswegen an dieser Stelle nicht weiter vertieft.

Nach der sukzessiven Analyse der Einzelsätze des aktuellen Eingabetextes mit Resolution textueller Bezüge erzeugt das Textverstehenssystem SYNDIKATE eine zu diesem Text gehörige Textwissensbasis, die dessen Inhalt entsprechend der erzielten Verstehenstiefe

widerspiegelt (Abb. 3.9, TKB_i).

An dieser Stelle sei noch abschließend angegeben, welchen quantitativen Umfang die einzelnen Wissensquellen von SYNDIKATE besitzen. In den Wissensbasen finden sich ca. 5.000 Konzepte und Relationen, von denen sich über 2.000 im domänenunabhängigen UPPER-MODEL befinden, je ungefähr 1.500 implementieren die Medizin- bzw. die Informationstechnologie-KB. Die verwendeten Lexika verfügen zusammengekommen über ca. 9.000 Einträge, von denen mehr als die Hälfte in den domänenunabhängigen Lexika für offene und geschlossene Wortklassen kodiert sind. Mit diesen Größenordnungen hat SYNDIKATE eindeutig den Rahmen eines “Spielzeugsystems” gesprengt. Allerdings steht eine Skalierung des Systems noch aus, die den momentan noch engen Domänenfokus auf Hardware in der IT und gastrointestinale Erkrankungen in der Medizin substantiell erweitert.

3.2.5 Textuelle Phänomene und Anaphernresolution

Eine unverbundene Analyse von Einzelsätzen reicht nicht aus, um für einen Eingabetext eine adäquate inhaltliche Repräsentation zu erzielen. Vielmehr basiert diese auf der Auflösung textueller Phänomene, die thematische Bezüge zwischen den Sätzen herstellen. Die inhaltliche Kohärenz eines Textes (wie auch jeden Diskurses) leitet sich maßgeblich aus der Verknüpfung der Teilaussagen zu einem Aussageganzen ab. Die Nicht-Beachtung textueller Bezüge führt bei der automatischen Textanalyse zum Aufbau von *unvollständigen*, *invaliden* und schließlich *inkohärenten* Textwissensbasen (Hahn et al., 1999a).

- Unvollständige Textwissensbasen entstehen, wenn im Vortext enthaltene Diskursentitäten nicht in den nachfolgenden Text übernommen werden, obwohl sie durch eine Proform referenziert werden, und dadurch eine inhaltliche Leerstelle entsteht (bei pronominalen Anaphern).
- Invalide Textwissensbasen entstehen, wenn ein im Text enthaltener Referent dort durch mehrere lexikalische Oberflächen referenziert wird und dieser Bezug nicht resolved wird (bei nominalen Anaphern).
- Inkohärente Textwissensbasen entstehen, wenn relationale Bezüge zwischen Diskursobjekten im Text, die nur implizit gegeben sind, nicht formal rekonstruiert werden

	Pronominale Anapher	Nominale Anapher	Textuelle Ellipse	Metonymie
grammatikalische Constraints	Numerus Genus	Numerus –	– –	– –
konzeptuelle Constraints	–	IS-A	role path patterns	type coercion

Tabelle 3.5: Wohlgeformtheitskriterien für Textphänomene

und dadurch wichtige Verknüpfungen verloren gehen (bei textuellen Ellipsen und Metonymien).

In Tabelle 3.5 sind die für die genannten textuellen Phänomene relevanten Wohlgeformtheitskriterien bei der Resolution aufgeführt. Dabei sticht hervor, daß von der pronominalen Anapher über die nominale Anapher bis hin zur textuellen Ellipse und der Metonymie die grammatikalischen Constraints abnehmen, wohingegen die konzeptuellen Constraints zunehmen. Beispielsweise verlangt eine pronominale Anapher die Kongruenz in Numerus und Genus mit dem Antezedenten und beinhaltet keinerlei konzeptuelle Constraints, die sich direkt aus der Beziehung zwischen Antezedenten und anaphorischem Ausdruck ableiten lassen. Bei der Nominalen Anapher muß neben der grammatikalischen Kongruenz im Numerus der Typ des Antezedenten mit dem anaphorischen Ausdruck in einer IS-A Beziehung, also einer Generalisierungsbeziehung stehen. Textuelle Ellipsen werden allein durch die Suche nach typischen Pfadmustern zwischen den Konzepten im Weltwissen resolved, durch die der Antezedent und der anaphorische Ausdruck miteinander relationiert werden. Grammatikalische Kriterien spielen hierbei indessen keinerlei Rolle mehr (Hahn et al., 1996b). Schließlich erfordert die Auflösung einer Metonymie auf der konzeptuellen Seite einen ausgefeilten Interpretation-Mechanismus (Markert & Hahn, 1997), der hier nicht weiter beschrieben wird.

Im folgenden soll nun die Auflösung der eingeführten textuellen Phänomene für ein konstruiertes Fragment aus authentischen Sätzen beispielhaft veranschaulicht werden,¹¹ denn einerseits liegt gerade in der Erweiterung des Satzverstehens um lokale textuelle

¹¹Das Beispiel setzt auf authentischem Textmaterial auf, reduziert dieses jedoch auf die hier relevanten Merkmale und kombiniert die entstehenden Teilsätze neu, um die für das Beispiel notwendige Kompaktheit zu gewährleisten.

Bezüge eine Besonderheit von SYNDIKATE. Andererseits lassen sich die in Kapitel 4.1 zur Darstellung kommenden Methoden zur semantischen Interpretation auf die Resolution intra- wie extrasententieller anaphorischer Phänomene ausdehnen. D.h., die zugrundeliegenden Schlußfolgerungs-Mechanismen sind abstrakt und flexibel genug, um auch die Berechnung referentieller Beziehungen zu unterstützen. Um die von SYNDIKATE behandelten Textphänomene zu erläutern und um aufzuzeigen, wie dabei vollständige, valide und kohärente Inhaltsrepräsentationen entstehen, soll hierzu der Satz (1), welcher bereits in Kap. 3.2.1 verwendet wurde, durch die nachfolgenden vier Sätze fortgeführt werden.

1. In einem reiskorngroßen Partikel mit einem Durchmesser von 3 mm wurde eine Magenschleimhaut vom Antrumtyp erfaßt.
2. Sie zeigt ein ödematöses Stroma.
3. Die Schleimhaut wird zudem dicht von Lymphozyten infiltriert.
4. Im Oberflächenschleim konnten *Helicobacter-pylori* nachgewiesen werden.
5. Der Patient muß weiterhin bioptisch kontrolliert werden.

Die jeweils im Diskurs textuell miteinander relationierten Wörter sind durch Unterstreichung gekennzeichnet. Um aber diese Diskursobjekte im jeweiligen textuellen Kontext verfügbar zu machen, benötigt das Textverstehenssystem ein Diskursgedächtnis, in dem diese gespeichert werden. In SYNDIKATE wird dieses Wissen über referenzrelevante Diskursobjekte auf der Grundlage des Centering-Ansatzes bereitgestellt (Grosz et al., 1995; Strube & Hahn, 1996). Das Centering modelliert die lokale Ebene eines kohärenten Diskurses, d.h. die Kohärenz zwischen einer Äußerung U_i und der darauffolgenden Äußerung U_{i+1} innerhalb eines Diskurssegments (z.B. eines Absatzes). Nicht in SYNDIKATE integriert sind Vorschläge zur Analyse globaler Kohärenz, d.h. der Kohärenz zwischen größeren Diskurseinheiten (vgl. hierzu Hahn & Strube (1997)). Diskurseinheiten, die der inhaltlichen Verknüpfung von Sätzen dienen, werden in sogenannten Centering-Listen organisiert. Jeder Äußerung U_i innerhalb eines Diskursabschnitts wird eine Menge an *forward-looking centers* $C_f(U_i)$ und ein einziger *backward-looking center*, $C_b(U_i)$, zugeordnet. Diese Centering-Listen bestehen aus partiell geordneten Tupeln. Die Ordnung der forward-looking centers $C_f(U_i)$ basiert nur auf der aktuell analysierten Äußerung U_i , wohingegen das backward-looking

(1)	Cb: — Cf: [PARTICLE.4-02: Partikel, DEGREE.10-06: Durchmesser, GASTRIC-MUCOSA.13-08: Magenschleimhaut, ANTRUM-TYPE.15-10: Antrumtyp]
-----	---

Tabelle 3.6: Centering-Daten für Satz (1) aus dem Textfragment

center des Folgesatzes, $C_b(U_{i+1})$, dasjenige Element aus $C_f(U_i)$ realisiert, für welches in U_{i+1} eine gültige Interpretation gefunden wurde. Jedes der in den centers enthaltenen Tupel besteht aus der Wortoberfläche des dazugehörigen Nomens im Text sowie der damit verbundenen Instanz im Weltwissen. Je weiter oben ein Tupel in dieser Liste steht, umso größer ist die Präferenz für dieses Listenelement als potentieller Antezedent aus U_i einer aufzulösenden Referenz in U_{i+1} . Die Ordnung der Liste erfolgt nach funktionalen Kriterien, die über reine Reihenfolgebedingungen oder die grammatikalische Funktion im Satz hinausgehen und sich auf topic/comment- und Thema/Rhema-Strukturen konzentrieren (Strube & Hahn, 1999).

Weil der Satz (1) außer der Reihenfolge der Nomina im Satz keine weiteren Ordnungskriterien enthält, wird die Centering-Liste für diesen Satz entsprechend dieser Reihenfolge aufgebaut. Sie ist in Tabelle 3.6 aufgeführt und besteht aus den jeweiligen Tupeln mit der Wortoberfläche im Satz und dem Instanzidentifizier aus der Wissensbasis (z.B dem Tupel PARTICLE.4-02: Partikel). Ein backward-looking center kann wegen des fehlenden Vortexes nicht angegeben werden.

3.2.5.1 Pronominale Anapher

Im Gegensatz zu den anderen im folgenden untersuchten Textphänomenen kann bei der pronominalen Anapher das Vorliegen eines anaphorischen Ausdrucks aus der Wortklasse abgeleitet werden. In SYNDIKATE wird die Resolution der Anapher ausgelöst, sobald der anaphorische Ausdruck, hier das Pronomen, seinen syntaktischen Kopf ermittelt hat. Jedoch treten pronominale Anaphern sowohl satzintern wie satzextern auf. Zuerst findet daher eine Überprüfung statt, ob die Referenz satzintern aufgelöst werden kann. An dieser Stelle greift natürlich nicht der Centering-Mechanismus, welcher für satzübergreifende Anaphern konzipiert ist. Vielmehr müssen im Dependenzgraphen für den aktuellen Satz als *D-Binding* beschriebene Abhängigkeitsmuster erfüllt sein (Strube & Hahn, 1995a).

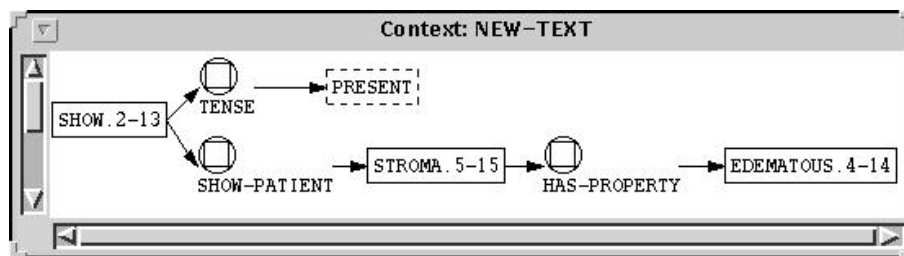


Abbildung 3.10: Satz (2) ohne Resolution der pronominalen Anapher

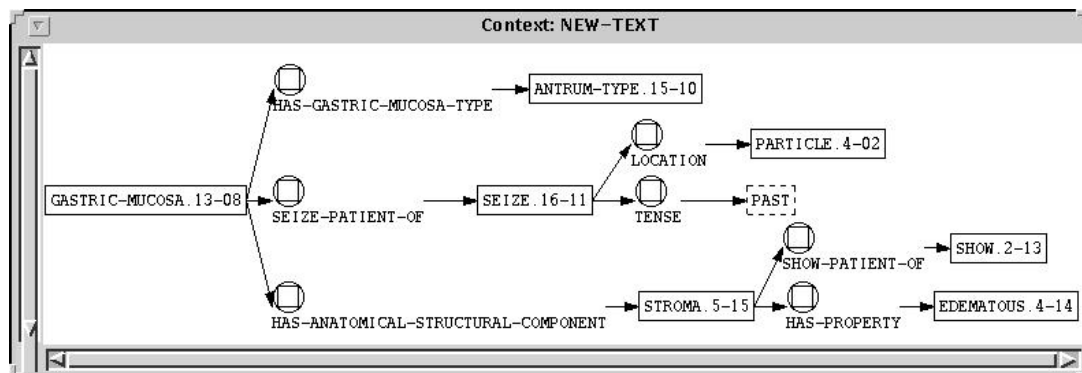


Abbildung 3.11: Sätze (1) und (2) mit resolvierter pronominaler Anapher

Abb. 3.10 zeigt den *unvollständigen* Konzeptgraphen für den Satz (2).¹² Die Unvollständigkeit resultiert aus der Tatsache, daß der Antezedent für die pronominalen Anapher GASTRIC-MUCOSA.13-08 nicht in der Bedeutungsrepräsentation des Satzes enthalten ist. Dadurch stehen die Sätze (1) und (2) unverbunden nebeneinander, obwohl sie referentiell verknüpft sind. Die Auflösung der pronominalen Anapher erfolgt, sobald der Wortaktor für „sie“ seinen lexikalischen Kopf „zeigt“ gefunden hat, um mit diesem eine Dependenzbeziehung einzugehen. Nun greift der Mechanismus zur Anaphernresolution auf die Centering-Liste des Vorsatzes zu. Der Wortaktor für „sie“ besitzt zwei disjunkte morphologische Merkmalslisten, i.e. (a.) (genus: fem; num: sing) und (b.) (genus: fem, mas, ntr; num: pl). Entsprechend den aus Tab. 3.5 ableitbaren Wohlgeformtheitskriterien muß eine Kongruenz mit dem Antezedenten in Numerus und Genus gegeben sein. Unter dieser Beschränkung wird nun die Centering-Liste des Vorsatzes durchgegangen. Erst das Tupel (GASTRIC-

¹²Auf eine grafische Präsentation der Dependenzgraphen wird in diesem Abschnitt verzichtet, weil diese keinen Beitrag zur Argumentation leisten.

(2)	Cb: GASTRIC-MUCOSA.13-08: Sie Cf: [GASTRIC-MUCOSA.13-08: Sie, STROMA.5-15: Stroma]
-----	---

Tabelle 3.7: Centering-Daten für Satz (2) aus dem Textfragment

MUCOSA.13-08, Magenschleimhaut) erfüllt diese Beschränkung, da weder Partikel noch Durchmesser das richtige Genus aufweisen. Es wird nach erfolgreicher Resolution aus dieser Liste gelöscht. Damit kann die Etablierung der Abhängigkeitsrelation zwischen *“zeigt”* und *“sie”* vorgenommen und entsprechend GASTRIC-MUCOSA.13-08 als Rollenfüller in der Instanz SHOW.2-13 eingetragen werden. Die abschließende konzeptuelle Interpretation führt zur Assertion GASTRIC-MUCOSA.13-08 HAS-ANATOMICAL-STRUCTURE STROMA.5-15 (siehe Abb. 3.11), wodurch die um das Thema *Magenschleimhaut* zentrierte Textpassage in textueller Hinsicht vollständig interpretiert ist.

Mit Konsumierung des Punktes als Indikator für das Satzende und dem Abschluß aller syntaktischen und semantischen Berechnungen wird die neue Centering-Liste für den Satz (2) erstellt (vgl. Tabelle 3.7). Die erfolgreich resolveerte Anapher kommt als einziges Element in das backward-looking center. Zugleich erfolgt die Ordnung der forward-looking centers nach funktionalen Kriterien, wonach ein kontextuell gebundenes Element das am stärksten präferierte Element für den Folgesatz ist. Dadurch rückt das Tupel (GASTRIC-MUCOSA.13-08, Sie) an die erste Position.

3.2.5.2 Nominale Anapher

Verglichen mit pronominalen Anaphern relaxieren nominale Anaphern die grammatikalischen Wohlgeformtheitsbedingungen zur Kongruenz im Numerus. Jedoch kommt als zusätzlicher konzeptueller Constraint die Subsumptionbeziehung zwischen anaphorischem Ausdruck und dem Antezedenten hinzu (vgl. Tabelle 3.5), d.h., der anaphorische Ausdruck ist konzeptuell betrachtet allgemeiner als der Antezedent. Da im Prinzip jedes Nomen einen anaphorischen Bezug zum Vortext haben kann, ist die Wortkategorie ein zu weites Triggerkriterium für das Auslösen der Anaphernresolution. Sie erfolgt daher unter der Beschränkung, daß ein Nomen sich in einer definiten Nominalphrase befindet und versucht, sich an seinen syntaktischen Kopf anzubinden.¹³ Die Suche nach dem Antezedenten erfolgt

¹³Diese Strategie wurde erfolgreich für Texte aus der Informationstechnologiedomäne in ITSyNDIKATE angewandt. Leider weisen die in der Medizin angetroffenen Texte zumeist den typischen Telegrammstil

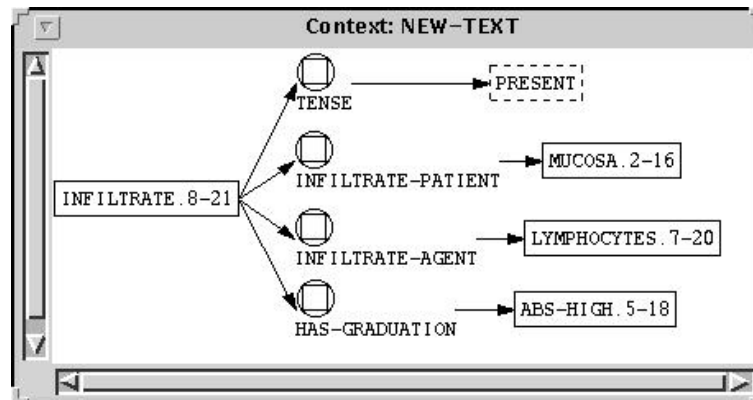


Abbildung 3.12: Satz (3) ohne Resolution der nominalen Anapher

unter Berücksichtigung der Constraints wie bereits oben beschrieben.

Die Betrachtung von Satz (3) und des dazugehörigen Konzeptgraphen (Abb. 3.12) zeigt eine vollständige konzeptuelle Interpretation der am Satz beteiligten Diskursobjekte INFILTRATE.8-21, MUCOSA.2-16, LYMPHOCYTES.7-20 UND ABS-HIGH.5-18. Nur bringt diese Form der Interpretation eine invalide Repräsentation des Textfragments mit sich, da die referentielle Identität der „Schleimhaut“ mit der „Magenschleimhaut“ aus dem Vortext nicht adäquat abgebildet ist. Veranschaulicht werden kann dies durch die nachfolgende Anfrage Q an die Textwissensbasis.¹⁴

Q : (retrieve ?x (Mucosa ?x))

A-: (|I| Mucosa.2-16, |I| Gastric-Mucosa.13-08)

A+: (|I| Gastric-Mucosa.13-08)

Anstatt nur einen Referenten wie in A+ wiederzugeben, resultiert die Anfrage in A- fälschlicherweise in der Ausgabe zweier Diskursobjekte.

Das richtige Ergebnis auf die Anfrage, i.e. (A+), entsteht durch die Auflösung der Nominalen Anapher. Sobald die definite Nominalphrase „die Schleimhaut“ ihren syntaktischen Kopf „wird“ gefunden hat, erfolgt das Durchlaufen der forward-looking centers des Satzes (2) (siehe Tabelle 3.7). Dies mündet in der Anfrage, ob eine GASTRIC-MUCOSA von MUCOSA subsumiert wird. Da dies offensichtlich erfüllt ist, werden die Konzeptgraphen der Sätze (1-2) und des Satzes (3) vereinigt, wobei die speziellere Instanz –in unserem Fall

auf, der sich üblicherweise auch in der Omission des definiten Artikels manifestiert.

¹⁴Die Anfrage ist hier in Form der Anfragesprache des Wissensrepräsentationssystems LOOM gegeben.

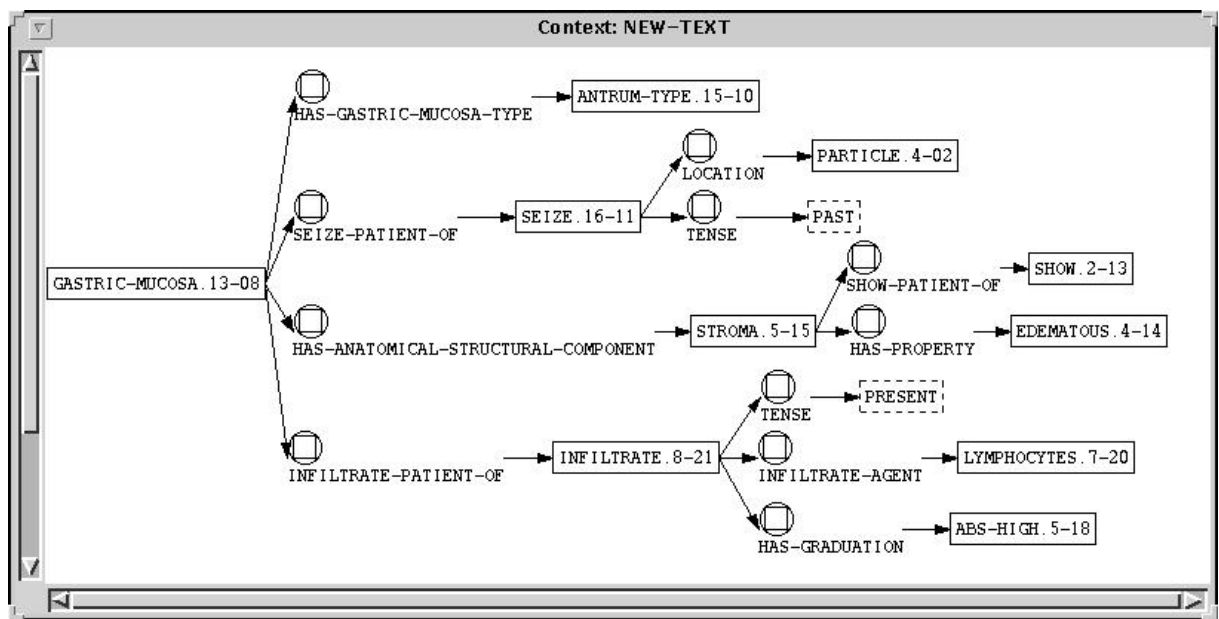


Abbildung 3.13: Sätze (1)-(3) mit resolvierter nominaler Anapher

(3)	Cb: GASTRIC-MUCOSA.13-08: Schleimhaut Cf: [GASTRIC-MUCOSA.13-08: Schleimhaut, LYMPHOCYTES.6-21: Lymphozyten]
-----	--

Tabelle 3.8: Centering-Daten für Satz (3) aus dem Textfragment

GASTRIC-MUCOSA.13-08– als Referent beibehalten wird. Das Ergebnis führt zu einer validen und vollständigen Inhaltsrepräsentation, wie sie in Abb. 3.13 grafisch zur Darstellung kommt. Hier ist GASTRIC-MUCOSA.13-08 anstelle von MUCOSA.2-16 als Füller an der Relation INFLTRATE-PATIENT-OF eingetragen. Am Satzende wird auch die Centering-Liste für den Satz (3) entsprechend dem Satzinhalt und des anaphorischen Bezugs zum Vortext aktualisiert (siehe Tabelle 3.8).

3.2.5.3 Textuelle Ellipse

Im Gegensatz zu den bisher eingeführten Textphänomenen weist die textuelle Ellipse keine grammatikalische Constraints bezüglich des Vortextes auf. Statt dessen relationiert die textuelle Ellipse einen quasi-anaphorischen Ausdruck mit seinem satzexternen Antezedenten. Dieser implizite Zusammenhang läßt sich zwar über die konzeptuelle Beschreibung

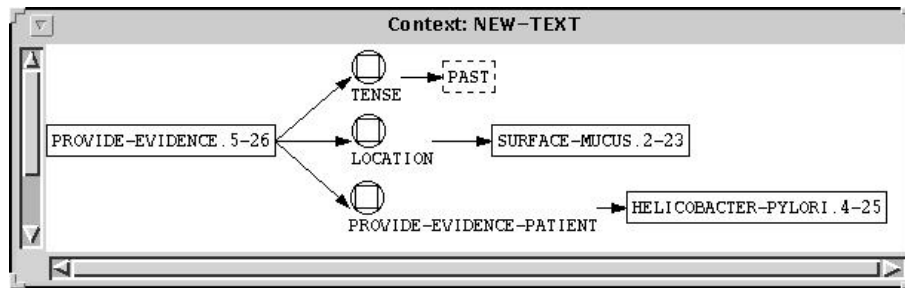


Abbildung 3.14: Satz (4) ohne Resolution der textuellen Ellipse

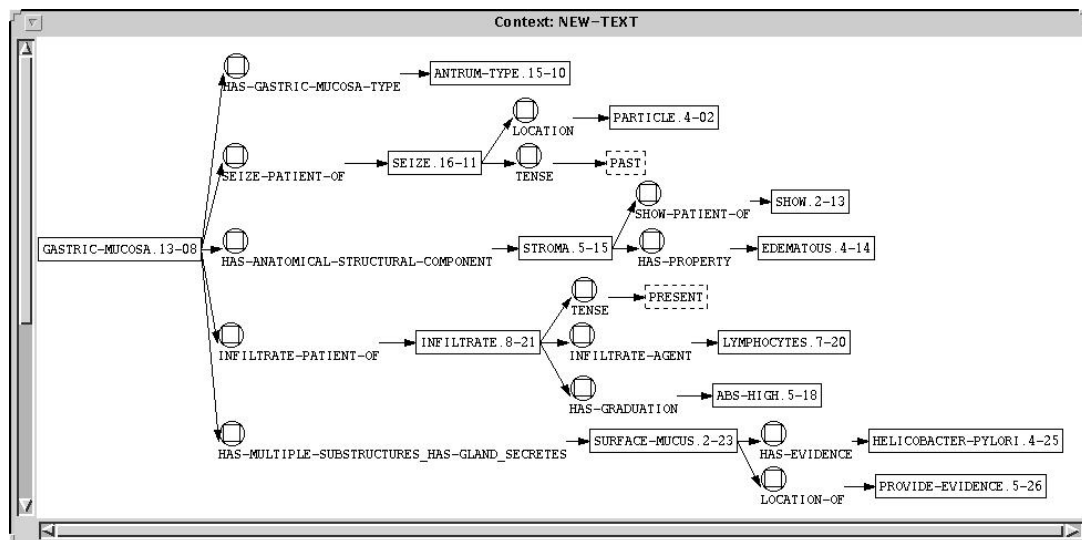


Abbildung 3.15: Sätze (1)-(4) mit resolvierter textueller Ellipse

der involvierten Referenten formal rekonstruieren, auf der Sprachoberfläche ist er hingegen zunächst nicht explizit. Da die textuelle Ellipse ebenfalls an das Auftreten in einer definiten Nominalphrase gebunden ist, komplementiert sie die nominale Anapher: Während die nominale Anapher eine Generalisierungsbeziehung zu ihrem Antezedent erfordert, verlangt die textuelle Ellipse die Rekonstruktion einer nicht-generalisierten konzeptuellen Relationierung. Die textuelle Kohärenz hängt folglich entscheidend von der Berechnung der richtigen Relationierung zwischen Antezedent und dem anaphorischen Ausdruck ab. An dieser Stelle spielt die Qualität und die Tiefe des verfügbaren Wissens eine entscheidende Rolle. Der zur Bestimmung möglicher Relationierungen zur Anwendung kommende Pfadsuchmechanismus wird in breiterem Kontext in Kapitel 4 erörtert. Bei dieser Suche ist eine mögliche Relationierung nicht auf direkt erreichbare Rollenfüller beschränkt, d.h.,

(4)	Cb: GASTRIC-MUCOSA.13-08: — Cf: [GASTRIC-MUCOSA.13-08: —, SURFACE-MUCUS.2-23: Oberflächenschleim, HELICOBACTER-PYLORI.4-25: Helicobacter-pylori]
-----	---

Tabelle 3.9: Centering-Daten für Satz (4) aus dem Textfragment

die Länge eines Pfads kann größer als 1 sein (Hahn et al., 1996b).

Im Satz (4) führt die definite Nominalphrase *“im Oberflächenschleim”* die textuelle Ellipse ein.¹⁵ Wegen einer Explizierung der textuellen Ellipse durch ein nachfolgendes Genitiv-Attribut (hier: *“im Oberflächenschleim der Magenschleimhaut”*) kann aufgrund der Inkrementalität des Parsing-Verfahrens die Resolution nicht erfolgen, wenn die definite Nominalphrase ihren syntaktischen Kopf gefunden hat. Weil sie den eigentliche Fortgang der inkrementellen Analyse ohnehin nicht beeinflusst, wird erst am Satzende versucht, die Relationierung des anaphorischen Ausdrucks zum Antezedenten zu bestimmen. Hierzu wird die Centering-Liste von Satz (3) genommen und die (noch) vorhandenen Tupel der Reihe nach auf ihre Relationierbarkeit mit der Instanz für *“Oberflächenschleim”* SURFACE-MUCUS.2-23 geprüft. Da dem Weltwissen zu entnehmen ist, daß eine Magenschleimhaut aus (anatomischen) Substrukturen besteht (HAS-MULTIPLE-SUBSTRUCTURES), zu denen Drüsen (GLAND) als anatomisches Teil gehören, welche wiederum Schleim (MUCUS) sekretieren, kann die in Abb. 3.15 enthaltene zusammengesetzte Relation HAS-MULTIPLE-SUBSTRUCTURES_HAS-GLAND_SECRETES eingezogen werden. Statt nun wie in Abb. 3.14 einen isolierten Teilsatz zu haben, beinhaltet Abb. 3.15 den impliziten Zusammenhang zwischen den Sätzen (1-3) und dem Satz (4). Der mit resolvierten Textphänomenen gewonnene Konzeptgraph ist damit kohärent, vollständig und valide. Die Tabelle 3.9 gibt nochmals die Centering-Daten nach Analyse des Satzes (4) wieder. Zu beachten ist, daß zwar der Identifier für den Referenten, GASTRIC-MUCOSA.13-08, im Weltwissen präsent ist, eine sprachliche Oberfläche findet sich wegen der Implizitheit des textuellen Bezugs freilich nicht.

¹⁵Der definite Artikel *“dem”* ist in diesem Beispiel mit der Präposition *“in”* zur Form *“im”* zusammengezogen.

3.2.5.4 Metonymie

Metonymien erweitern die für textuelle Ellipsen typischen Sprachmuster um bildliche Sprache (Markert & Hahn, 1997). Ebenso wie textuelle Ellipsen stellen sie keine Kongruenzbedingungen bezüglich des Antezedenten. Statt dessen werden noch komplexere Berechnungen auf der konzeptuellen Seite erforderlich. Eine Äußerung A wird als metonymisch angesehen, wenn sie von ihrer Standarddenotation abweicht, indem sie für eine andere Entität B steht, welche nicht explizit im aktuellen Kontext enthalten ist. Jedoch ist B mit A über eine normalerweise konventionalisierte konzeptuelle Relation r verbunden. Das Vorhandensein eines Typenkonflikts stellt oftmals einen Indikator für das Vorliegen einer Metonymie dar. Der metonymische Ausdruck in Form eines Nomens oder einer Nominalphrase steht zu einem Antezedenten außerhalb des aktuellen Satzes in Beziehung und bringt typisierbare Wohlgeformtheitsbedingungen mit sich, wie der metonymische Ausdruck und der Antezedent miteinander relationiert werden können. Ebenso wie bei textuellen Ellipsen muß die fehlende konzeptuelle Relationierung anhand des Domänenwissens und der Idiosynkrasien bei der Verwendung von Metonymien berechnet werden. Wie bei der nominalen Anapher besteht die Triggerbedingung bei der Metonymieberechnung aus der Präsenz einer definiten Nominalphrase, die sich an ihren syntaktischen Kopf anzubinden versucht. Bei der Auflösung der Metonymie gibt es keine Präferenz für die wörtliche Bedeutung. Vielmehr behandelt der Pfadsuch-Algorithmus wörtliche und figurative Sprache im Unterschied zu anderen Ansätzen gleichwertig, d.h., die Berechnung erfolgt gleichzeitig. Für die Auswahl der richtigen Relationierung existieren konzeptuelle Plausibilitätskriterien, welche in einer Selektionheuristik zur Anwendung kommen (Hahn & Markert, 1997). Damit wird ein exorbitantes Anwachsen von Lesarten wie z. B. in Bouaud et al. (1996) verhindert.

Im Satz (5) unseres Beispiels wird die Prüfung nach konzeptuellen Wohlgeformtheitskriterien ausgeführt, sobald die definite Nominalphrase *“der Patient”* versucht, mit dem syntaktischen Kopf *“muß”* eine Dependenzrelation einzugehen. Mit der syntaktischen Anbindung des Teilgraphen (*“werden”* – *vrupart* – *“kontrolliert”*) an den syntaktischen Kopf *“muß”* werden die konzeptuellen Wohlgeformtheitsbedingungen des Vollverbs *“kontrolliert”* verfügbar. Im Weltwissen erlaubt die konzeptuelle Beschränkung laut Definition am Konzept CHECK, dem konzeptuellen Korrelat von *“kontrolliert”*, keinen Füller vom Typ PATIENT.¹⁶ Der dadurch entstehende Typenkonflikt untersagt die Etablierung der Depen-

¹⁶Das Konzept PATIENT modelliert eine Person, welche sich einer medizinischen Untersuchung oder

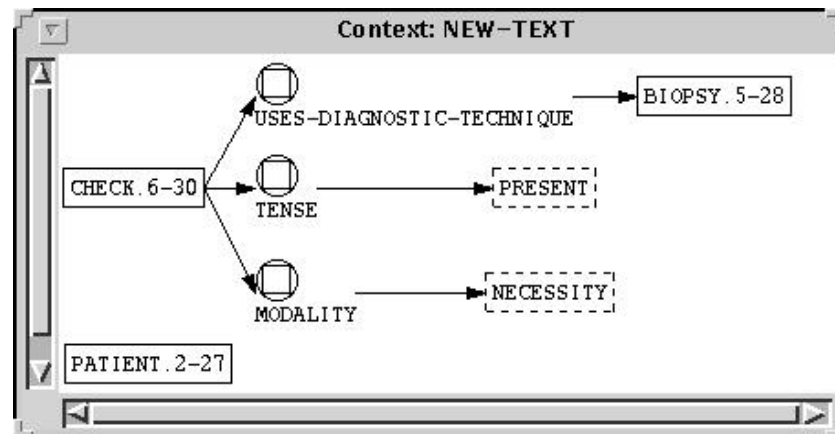


Abbildung 3.16: Satz (5) ohne Resolution der Metonymie

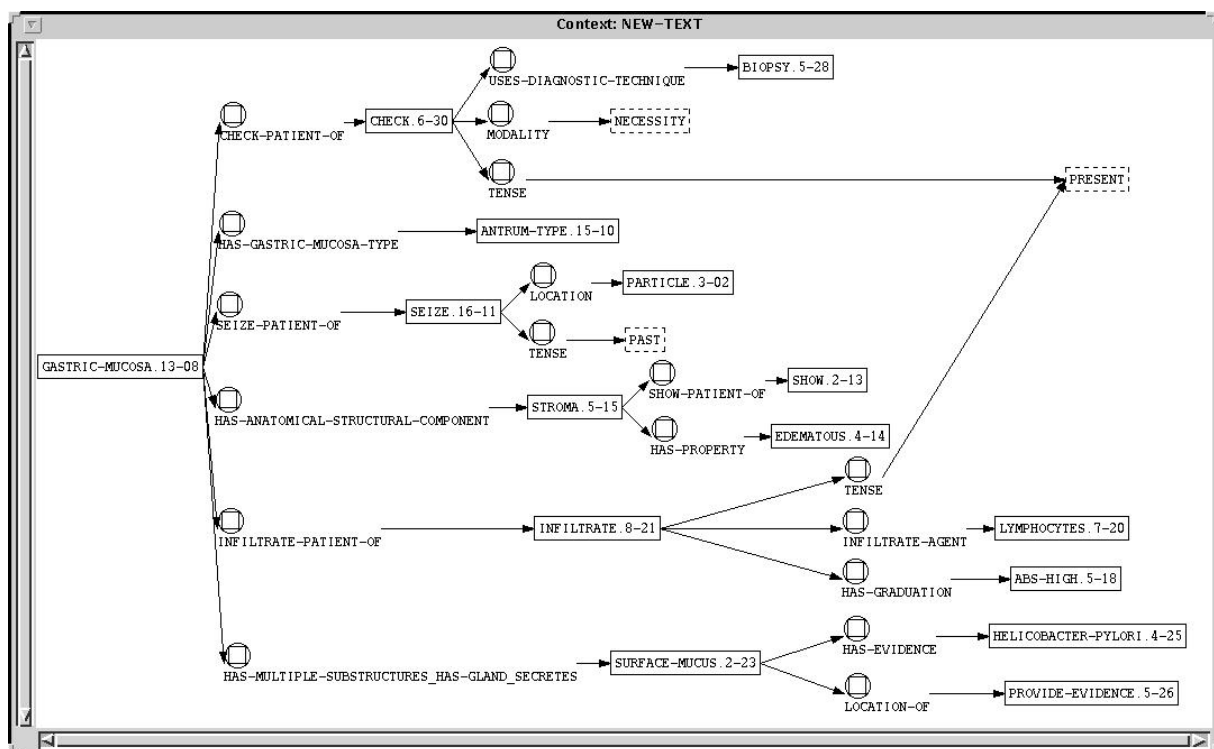


Abbildung 3.17: Sätze (1)-(5) mit Resolution der Metonymie

denzrelation im Fall einer wörtlichen Interpretation von PATIENT. Die Ergebnisse der Metonymieauflösungen erlauben jedoch die Anbindung, da unter Verwendung der Centering-Daten aus dem vorangegangenen Satz (4) eine gültige Lesart erzeugt werden kann. Für das erste Element der Centering-Liste, GASTRIC-MUCOSA.13.08, wird überprüft, ob sich ein gültiger metonymischer Pfad zu PATIENT.2-27 finden läßt. In unserem Beispiel liegt eine – übrigens stark konventionalisierte – *Ganzes-für-Teil* Metonymie vor. Über die HAS-ANATOMICAL-PART-Relation, einer Unterrelation von HAS-PART, lassen sich PATIENT und GASTRIC-MUCOSA relationieren. GASTRIC-MUCOSA wiederum ist als Unterkonzept von ORGANISM-SUBSTRUCTURE ein legaler Füller am Konzept CHECK innerhalb der medizinischen Domänenwissensbasis. Das Einziehen der entsprechenden Relationierungen ergibt den Abb. 3.17 zu entnehmenden Konzeptgraphen. Im Gegensatz zur Einzelsatzanalyse aus Abb. 3.16 ist GASTRIC-MUCOSA.13-08 als Referent der ärztlichen Handlung CHECK.6-30 eingetragen. Weiterhin erfolgt die Einbindung der Diskursentität PATIENT.2-27 als metonymischer Ausdruck, die in Abb. 3.16 abseits steht, durch die Relationierung mit GASTRIC-MUCOSA.13-08 über die entsprechende HAS-ANATOMICAL-PART Rolle. Dies ist aus Gründen der Übersichtlichkeit in Abb. 3.17 jedoch nicht mehr dargestellt.

Stellt man die Ergebnisse der fünf Einzelanalysen dem um das Thema GASTRIC-MUCOSA.13.-08 zentrierten Konzeptgraphen in Abb. 3.17 gegenüber, so ist ohne weiteres ersichtlich, daß der Verlust an Textaussagen bei unterlassener Resolution von Textphänomenen die Kohärenz eines Diskurses aushebelt, ja zum Teil sogar zu falschen inhaltlichen Repräsentationen führt. Nur die kontinuierliche Einbindung anaphorischer Referenzbeziehungen in den Textanalyseprozeß läßt eine vollständige, valide und kohärente Textwissensbasis entstehen. Hierzu werden die Einzelaussagen fortlaufend an den Diskursreferenten im thematischen Fokus akkumuliert.

3.2.5.5 Empirische Validierung zu Textphänomenen

Zweifelsohne stellen Textphänomene einen verstehensrelevanten Bestandteil jeglichen Textes dar.¹⁷ Jedoch divergiert im allgemeinen die Häufigkeit der verwendeten Anaphern sowie die Verteilung der Textphänomene relativ zueinander. Strube (1997) untersuchte im Rah-

Behandlung unterzieht.

¹⁷Andere wichtige Mittel textueller Verknüpfung wie z. B. Deixis werden in dieser Passage nicht berücksichtigt, auch deshalb, weil sie sich grundsätzlich nicht auf die hier beschriebene *lokale* Kohärenz beschränken lassen.

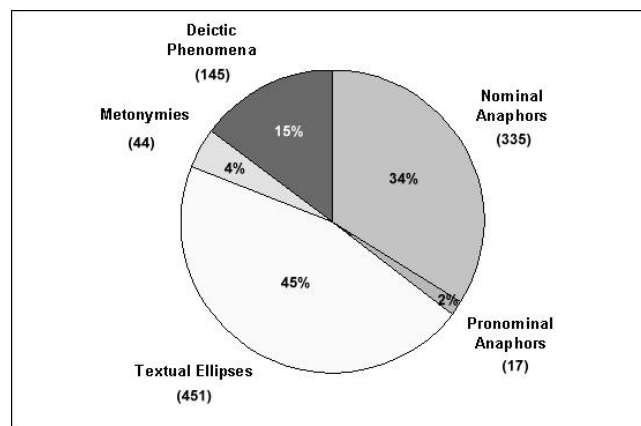


Abbildung 3.18: Distribution von Textphänomenen in medizinischen Texten

men von ITSYNDIKATE die Verteilung und Häufigkeit von Anaphern in Abhängigkeit von der Textsorte bezüglich literarischer Texte, Artikel in politischen Zeitschriften sowie von Hardwaretestberichten aus Computer-Fachzeitschriften. Offensichtlich interagiert der Gebrauch sowohl mit der Textsorte als auch mit der Länge eines Texts. Um aber auch die Medizindomäne abzudecken und die Relevanz von Textphänomenen ebenfalls für diesen Bereich zu belegen, der stark von großen Mengen an Alltags- und Routinetexten geprägt ist, wurde eine weitere Studie für MEDSYNDIKATE erforderlich.

In diesem Zusammenhang konnte in einer empirischen Untersuchung an einem randomisierten Beispielsatz von Histologieberichten zu einem breiten Spektrum an Erkrankungen, die aus dem elektronischen Archiv der Pathologie des Universitätsklinikums Freiburg entstammen, nachgewiesen werden, daß textuelle Phänomene in der Medizin ebenfalls einen relevanten Faktor darstellen (Hahn et al., 1999a). Dazu wurde eine Stichprobe von 100 Texten aus dem Klinikinformationssystem extrahiert. Da die Gesamtzahl an Wörtern 14.000 betrug, ergab sich ein Durchschnitt von 140 Wörtern pro Text bei einem Minimum von 23 Wörtern und einem Maximum von 925 Wörtern. Überraschenderweise dominieren in Hinblick auf die quantitative Distribution der Phänomenklassen in der Medizin vor allem textuelle Ellipsen mit 45% [42%–48%]¹⁸ und nominale Anaphern mit 34% [31%–37%] (Abb. 3.18, dort auch die absoluten Zahlen). Pronominale Anaphern spielen mit einer Häufigkeit von 2% [1%–3%] so gut wie gar keine Rolle. Da insbesondere die Auflösung textueller Ellipsen wie auch die der nominalen Anaphern stark wissensbasierte formale Rekonstruktionen

¹⁸Bei allen Prozentangaben sind die Konfidenzintervalle jeweils in eckigen Klammern mit angegeben.

erfordert, bestätigt die Verteilung der Phänomene die Tatsache, daß medizinische Texte durch ihren hohen Grad an Impliztheit zumeist nur durch den vorgebildeten Experten zu verstehen sind. Gerade wegen der nicht nachzuvollziehenden impliziten Textbezüge bleiben diese Texte dem Nicht-Experten unzugänglich.

Kapitel 4

Semantische Interpretation

In diesem Kapitel stelle ich ein schlankes Modell zur semantischen Interpretation vor, daß auf wohldefinierten Konstellationen in Abhängigkeitsgraphen basiert. Dadurch, daß sich die semantische Interpretation aus einem einzigen generischen Interpretationsschema ableitet, von dem etwas mehr als zehn Spezialisierungen abgeleitet sind, gibt es nur eine extrem kleine Regelmenge. Diese enorme Ökonomie bei der Regelkodierung wird erreicht, indem zum einen die strenge Modularisierung der Wissensquellen ausgenutzt wird und zum anderen Spezifikationen auf allen Wissensebenen in abstrahierenden Vererbungshierarchien abgelegt sind.

Aus den dargestellten Merkmalen ergeben sich eine Reihe von Vorzügen. Die Regelmenge ist sehr begrenzt. Selbst Interpretationen komplexerer Sprachphänomene wie Präpositionalphrasen werden durch eine einzige parametrisierte Regel abgedeckt. Darüber hinaus führen Erweiterungen an den Wissensquellen, beispielsweise der Wissensbasis oder dem Lexikon, nicht dazu, daß eine weitere Interpretationsregel für Präpositionalphrasen erforderlich wird. Genau aus diesem Grund wird die Portierung auf andere Domänen erleichtert. Die IT und die Medizin nutzen dieselbe Regel zur Interpretation von Präpositionalphrasen. Grundsätzlich reicht diese Regel für jede weitere Domäne aus. Im Resultat gewährleistet das in SYNDIKATE eingesetzte schlanke Verfahren zur semantischen Interpretation eine einfache Wartbarkeit und Portierbarkeit. Da die semantischen Regeln nur lokal auf Teilgraphen des Abhängigkeitsgraphen angewendet werden und dabei auf genau zwei Konfigurationen aufsetzen, ist der Kontrollfluß der Regeln absolut transparent.

4.1 Ein schlankes Modell zur semantischen Interpretation

Die semantische Ebene in einem Textverstehenssystem bezeichnet im allgemeinen die Interpretationsschicht, die die Abbildung syntaktischer Strukturbeschreibungen, etwa in Form eines Phrasenstrukturbaums, auf eine inhaltliche Repräsentation leistet. Entsprechend den Entwurfsprinzipien von SYNDIKATE sind die Wissens Ebenen, aus deren Zusammenspiel eine inhaltliche Interpretation für einen Text gewonnen wird, voneinander getrennt. Auf der syntaktischen Ebene ist keinerlei semantisches oder konzeptuelles Wissen kodiert. Die Domänenwissensbasis ihrerseits abstrahiert vom eigentlichen Sprachgebrauch und stellt das Wissen in seinen ontologischen, weitestgehend sprachunabhängigen Zusammenhängen dar. Schließlich liegt das Wissen für die Transition von syntaktischen zu inhaltlichen Beschreibungen ausschließlich in der semantischen Ebene.

Eine Besonderheit von SYNDIKATE besteht darin, daß die semantischen Berechnungen unmittelbar auf der Textwissensbasis durchgeführt werden. Die diese semantischen Berechnungen verwaltenden semantischen Objekte sind in demselben Formalismus repräsentiert, in dem auch das Domänenwissen beschrieben ist. Dementsprechend gibt es keine intermediäre Repräsentationsschicht auf der semantischen Ebene.

Zur Einführung des in SYNDIKATE verwendeten Ansatzes zur semantischen Interpretation sei ein etwas komplexerer, zu Illustrationszwecken konstruierter Beispielsatz eingeführt, an dem der Ansatz zur semantischen Interpretation vom Prinzip her veranschaulicht werden kann.

6. Compaq verkauft Computer der Pentium-Klasse mit Festplatten, die von IBM hergestellt werden.

Angenommen, für die in Satz 6 enthaltenen Wörter läge systemseitig eine vollständige Spezifikation vor, dann können nur für eine Teilmenge der im Satz enthaltenen Wörter auch unter diesen Idealbedingungen konzeptuelle Korrelate angegeben werden. Im Beispielsatz sind dies *“Compaq”*, *“verkauft”*, *“Computer”*, *“Pentium-Klasse”*, *“Festplatten”*, *“IBM”* sowie *“hergestellt”*. Wörter, denen konzeptuelle Korrelate zugewiesen werden können, gehören generell zu den offenen Wortklassen und sind Inhaltswörter. Die verbleibende Restmenge besteht aus Wörtern (*“der”*, *“mit”*, *“die”*, *“von”* und *“werden”*), die über keine eigenwertige Bedeutung verfügen. Sie können daher inhaltlich nicht gefüllt werden. D.h.,

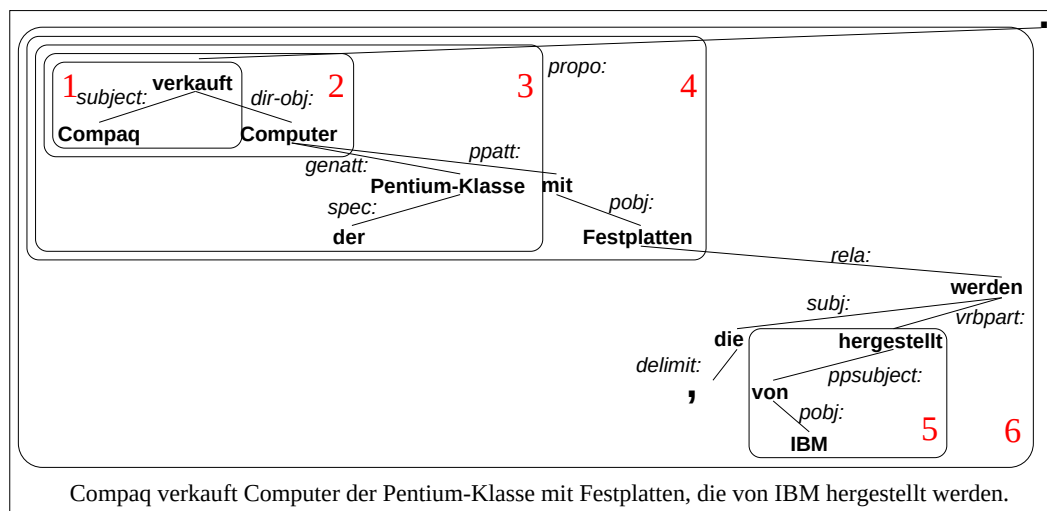


Abbildung 4.1: Dependenzgraph für Satz (6)

im ontologischen Sinne korrelieren sie nicht mit einer Entität irgendeines Diskursbereichs. Dennoch tragen sie entscheidend zu einer adäquaten inhaltlichen Interpretation eines Satzes bei, indem sie semantische Restriktionen für Relationierungen zwischen Inhaltswörtern mit sich bringen oder Skopus- und Quantifikationsinformationen tragen. Im allgemeinen können diese Wörter zu den geschlossenen Wortklassen gezählt werden.

Der innerhalb von SYNDIKATE realisierte Ansatz zur semantischen Interpretation setzt auf der Beschreibung einer begrenzten Menge von typischen Konstellationen in einem Dependenzgraphen auf, die Muster von Teilgraphen verwenden, um diese zunächst lokal und unabhängig vom Satzkontext zu interpretieren (womit der Bezug zur Robustheit des Parsingverfahrens unmittelbar deutlich wird). Darüber hinaus sind die Wissens Ebenen modularisiert und basieren auf kodierungsökonomischen Beschreibungen in Vererbungshierarchien. Der hierdurch erreichbare Grad an Abstraktion in der Beschreibung garantiert die notwendige Flexibilität und Verallgemeinerbarkeit des Ansatzes, die die Anwendung in heterogenen Domänen ermöglichen, ohne den Anpassungsaufwand exorbitant anwachsen zu lassen.

Der herausragendste Vorteil des in SYNDIKATE verwendeten Ansatzes zur semantischen Interpretation besteht, wie im folgenden gezeigt wird, darin, daß eine sehr kleine Zahl semantischer Regeln ausreicht, um eine große Menge von Interpretationen abzudecken, die im Laufe des Textverstehens zu verarbeiten sind.

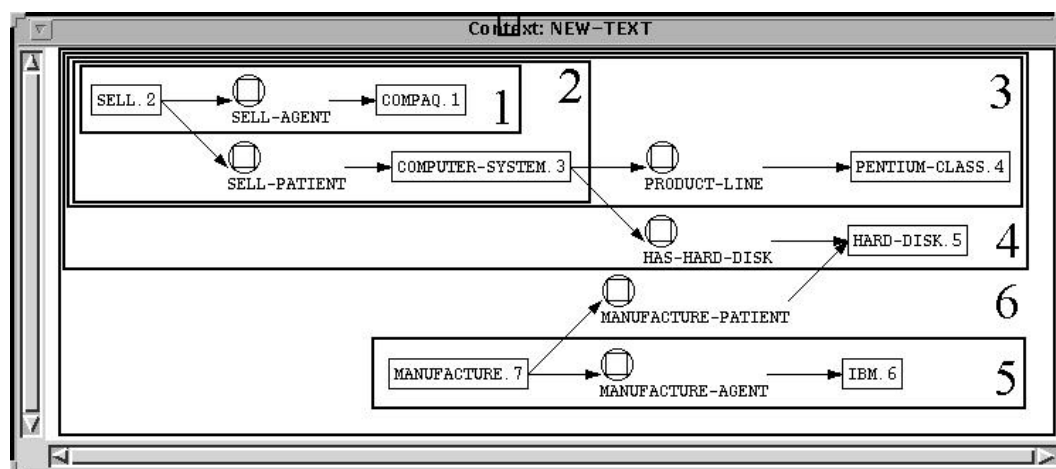


Abbildung 4.2: Konzeptgraph für Satz (6)

4.1.1 Semantisch interpretierbare Subgraphen

Ein Blick auf die in Abb. 4.1 dargestellte Dependenzstruktur für Satz (6) weist verschiedene Konfigurationen auf, in denen Tupel von Inhaltswörtern im Dependenzgraphen über sogenanntes *direct linkage* und *mediated linkage* miteinander verbunden sind. Aus diesen Konfigurationen leiten wir den Begriff eines *minimalen interpretierbaren Subgraphen* ab.

Direct Linkage. Zwei Inhaltswörter sind direkt miteinander verbunden, wenn die Pfadlänge des Kantenzugs an Dependenzrelationen zwischen ihnen genau 1 beträgt. In Abb. 4.1 ist dies beispielweise für die Teilgraphen (“Compaq” – subject – “verkauft”), (“verkauft” – direct-object – “Computer”) sowie (“Computer” – genitive-attribute “Pentium-Klasse”) der Fall.

Mediated Linkage. Zwei Inhaltswörter sind indirekt miteinander verbunden, wenn es einen sie verbindenden Kantenzug an Dependenzrelationen gibt, dessen Pfadlänge größer als 1 ist und auf dem keine Inhaltswörter traversiert werden. Dies gilt im vorliegenden Beispiel für die Kantenzüge in den entsprechenden Subgraphen (“Computer” – pparent – “mit” – pobject – “Festplatten”), (“hergestellt” – ppsubject – “von” – pobject – “IBM”) sowie (“die” [= Festplatten] – subject – “werden” – verbpart – “hergestellt”).¹

D.h., gemäß der Definitionen für direct und mediated linkage bildet der Kantenzug “verkauft” – direct-object – “Computer” – genitive-attribute – “Pentium-Klasse” keinen *mini-*

¹Der im letzten Kantenzug enthaltene Spezialfall einer satzinternen Proform wird in der nachfolgenden Erläuterung der verschiedenen Teilgraphen ausführlicher zur Darstellung kommen.

malen interpretierbaren Subgraphen, da der Pfad von “verkauft” nach “Pentium-Klasse” über das Inhaltswort “Computer” führt.

Werden während der inkrementellen syntaktischen Analyse zwei Inhaltswörter entweder durch *direct* oder *mediated linkage* miteinander verbunden, so bildet der entstehende Teilgraph im Dependenzgraphen einen minimalen interpretierbaren Subgraphen. Wann immer ein minimaler interpretierbarer Subgraph vollständig ist, erfolgt unmittelbar seine semantische Interpretation, d.h. die Berechnung einer konzeptuellen Relationierung zwischen den Inhaltswörtern, die in der Textwissensbasis die beiden Instanzen repräsentieren. Damit entspricht der sukzessiven Vereinigung von Teilgraphen auf der Parserseite über Dependenzrelationen die kompositionale Erweiterung der Interpretationsstrukturen in der Textwissensbasis durch konzeptuelle Relationen.

Der Vergleich von Abb. 4.1 mit Abb. 4.2 illustriert die skizzierten Zusammenhänge. Zunächst existiert für jedes der über *direct* oder *mediated linkage* syntaktisch relationierten Paare an Inhaltswörtern eine konzeptuelle Relationierung in der Wissensbasis. So kann der direkten Verbindung zwischen “Compaq” und “verkauft” über die *subject* Relation auf der Syntaxseite der Teilkonzeptgraph SELL.2 SELL-AGENT COMPAQ.1 zugeordnet werden. Die zweite Konfiguration eines minimalen interpretierbaren Subgraphen besteht aus einer *direct linkage*, die “Computer” über die Dependenzrelation *direct-object* an das finite Verb “verkauft” anbindet. Auf Seiten der Textwissensbasis berechnet die semantische Interpretation die Relation SELL-PATIENT, was in der Assertion SELL.2 SELL-PATIENT COMPUTER-SYSTEM.3 resultiert. Weiterhin läßt sich für die vermittelte Verbindung des Tupels (“Computer”, “Festplatten”) über den intervenierenden Knoten *mit* und die Kanten *pattribute* und *pobject* der konzeptuelle Teilgraph COMPUTER-SYSTEM.3 HAS-HARD-DISK HARD-DISK.5 berechnen.

In Abb. 4.1 markieren die Ziffern 1-6 die jeweils vollständigen semantisch interpretierbaren Subgraphen, wie sie sich während der inkrementellen Satzanalyse aufbauen. Jedem dieser durch die Ziffern markierten Zustände bei der Berechnung des Dependenzgraphen für den Gesamtsatz entspricht ein Konzeptgraph in Abb. 4.2, der die inhaltliche, fortlaufend kompositional erweiterte Repräsentation der jeweiligen Analysezustände reflektiert (ebenfalls mit 1-6 durchnummeriert). Der enge Zusammenhang zwischen Dependenzstrukturen und den diese repräsentierenden Konzeptgraphen wird damit evident (vgl. Hajicova (1987)). Den Aufsatzzpunkt für die enge Koppelung der Beschreibungsebenen für einen Eingabesatz stellt der minimale interpretierbare Subgraph im Dependenzgraphen dar.

Definition 4.1 Seien w und w' zwei Inhaltswörter und n die Anzahl der Wörter, die einen Subgraphen im Dependenzgraphen begrenzen. Darüber hinaus seien $w_2, \dots, w_{n-1} \in S$ mit $(n \geq 2)$ Präpositionen, Modalverben oder Auxiliärverben und $w_1 := w$ und $w_n := w'$. w und w' stehen in einer mediatisierten syntaktischen Relationierung zueinander genau dann, wenn es einen Index $l \in \{1, \dots, n\}$ gibt, so daß die folgenden Bedingungen gelten:

1. w_i ist Modifikator von w_{i+1} für $i \in \{1, \dots, l-1\}$;
2. w_i ist Kopf von w_{i+1} für $i \in \{l, \dots, n-1\}$.

Eine solche beliebige Folge $w_1, \dots, w_n$ bezeichnet einen minimalen semantisch interpretierbaren Subgraphen des Dependenzgraphen für den Satz S .

Definition 4.1 für die mediatisierte syntaktische Relationierung umfaßt zum einen die Umschreibung einer *direct linkage* ($n := 2$, so daß eine leere Menge von intervenierenden Knoten entsteht). Zum anderen beinhalten die Spezialfälle $l := 1$ und $l := n$ eine Menge von aufsteigenden bzw. absteigenden Kopf-Modifikator Relationierungen. Mit den eingeführten Fällen für *direct* und *mediated linkage*, welche in Definition 4.1 formal gefaßt sind, sind die beiden relevanten Konfigurationen innerhalb von Dependenzgraphen beschrieben. Im folgenden wird auf den semantischen Interpretationsprozeß selbst eingegangen.

Die Berechnung der konzeptuellen Relationen zwischen den Instanzen der Textwissensbasis verwendet einerseits Restriktionen, wie sie aus den Dependenzrelationen auf der syntaktischen Ebenen ableitbar sind. Andererseits beschränken auch die intervenierenden Knoten, sofern vorhanden, die semantische Interpretation. Die semantische Ebene verbindet das lexikalische und konzeptuelle Wissen durch abstrahierende Beschreibungen und weitestgehende Generalisierungen. Zunächst relationieren *statische* semantische Constraints grammatikalisches und konzeptuelles Wissen durch eine a priori fixierte Abbildungsoperation von Dependenzrelationen auf eine Menge von konzeptuellen Relationen, die auf beiden Beschreibungsebenen vom Vererbungsmechanismus Gebrauch macht. Anschließend werden diese statischen Wohlgeformtheitsbedingungen *dynamisch* durch eine Pfadsuche in dem das Weltwissen konstituierenden Konzeptgraphen expandiert, um den aktuellen Dependenzgraphen in seine entsprechende konzeptuelle Interpretation zu überführen.

Abb. 4.3 verdeutlicht den statischen Teil der Interpretation in seiner Verknüpfung mit den Vererbungshierarchien innerhalb der Wissensmodule. Auf der linken Seite ist eine Teilmenge der in PARSETALK verwendeten Dependenzrelationen $\delta \in D$ dargestellt.² Auf der

²Die eckigen Klammern erscheinen deshalb in der Grafik, weil in den Abbildungen von Dependenzgra-

rechten Seite ist ein Fragment der Wissensbasis dargestellt. Innerhalb der Wissensbasis gibt es zwei Hierarchien, eine Relationenhierarchie und eine Konzepthierarchie, die jeweils auf Generalisierungsbeziehungen basieren. In Abb. 4.3 ist dieser Zusammenhang durch die Labelung $isa_{\mathcal{R}}$ bzw. $isa_{\mathcal{F}}$ der gerichteten Kanten zwischen Relations- bzw. Konzeptbezeichnern verdeutlicht. Durch die Verwendung von Relationen zur Konzeptdefinition sind beide Hierarchien aufeinander bezogen. Zum Beispiel wird das Konzept `SELL` u.a. durch die Relation `SELL-AGENT` als Subrelation von `AGENT` definiert.

Die syntaktische und konzeptuelle Ebene stehen jedoch nicht isoliert nebeneinander. Ausgehend von dem Faktum, daß eine Teilmenge der Dependenzrelationen Restriktionen für die semantische Interpretation mit sich bringt, sind einige Dependenzrelationen statisch mit Top-Level-Relationen, d.h. den obersten Relationen der konzeptuellen Relationenhierarchie verdrahtet. Diese Dependenzrelationen erscheinen in der Abbildung kursiv. Beispielsweise muß die Dependenzrelation *direct-object* konzeptuell stets in Form einer `PATIENT-` oder `CO-PATIENT-` Relation interpretiert werden. Andere Dependenzrelationen, wie etwa *genitive-attribute* bergen keine solchen Einschränkungen, d.h., sie haben keinerlei Einfluß auf den semantischen Interpretationsprozeß. Sie sind daher auch nicht statisch mit der Relationenhierarchie verbunden.

Durch diese statische Kopplung zeigt die Syntax also nicht nur an, welche Entitäten innerhalb eines Textes konzeptuell zu relationieren sind, sondern sie bringt in unserem Ansatz auch noch Beschränkungen darüber mit sich, welche Relationierungen überhaupt erlaubt sind. Die Wohlgeformtheitsbedingungen, welche aus den grammatikalischen Dependenzbeziehungen abgeleitet werden können, sind in zwei unterschiedlichen Formen kodiert, nämlich einerseits als Positiv-Liste, D_+^{lexval} , und andererseits als Negativ-Liste, D_-^{lexval} . Aus diesen lassen sich die für die semantische Interpretation zulässigen Relationen, R_+ , wie auch die ausgeschlossenen konzeptuellen Relationen, R_- , ermitteln, so wie es durch das einfache statische Verbinden der syntaktischen Relationen mit den konzeptuellen festgeschrieben ist.

Das Wissen über D_+^{lexval} und D_-^{lexval} gehört zur Valenzspezifikation und ist auf der Ebene der Wortklassen $\mathcal{W}$ in die Klassenhierarchie integriert, so daß $lexval \in \mathcal{W} \times \mathcal{D}$.

phen, die die syntaktischen Strukturen zu Eingabesätzen darstellen, bei Bezeichnern für Dependenzrelationen aus Gründen der Übersichtlichkeit Abkürzungen benutzt werden. Die Teilstrings der Bezeichner ohne die Buchstabenfolge in den eckigen Klammern sind demnach die in den Abbildungen zu Dependenzgraphen anzutreffenden Bezeichner für Dependenzrelationen.

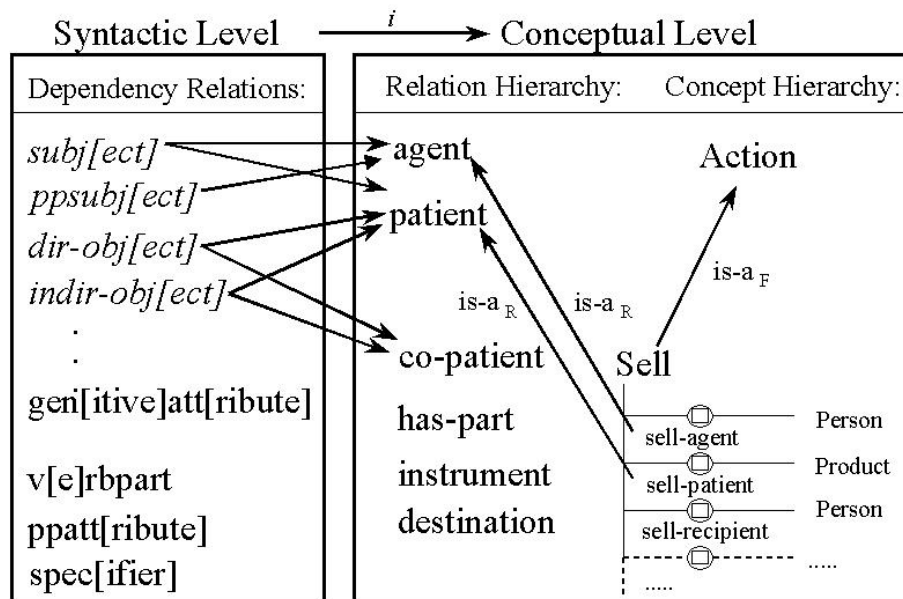


Abbildung 4.3: Synthese der Wissens Ebenen bei der statischen Interpretation

Damit erben alle zu einer Wortklasse gehörenden Lexeme diese Beschreibungen. Beispielsweise definiert die Wortklasse für transitive Verben, $\text{VERBTRANS} \in \mathcal{W}$, für die Subjektvalenz $D_+^{\langle \text{verbtrans}, \text{subject} \rangle} := \{\text{subject}\}$ und $D_-^{\langle \text{verbtrans}, \text{subject} \rangle} := \emptyset$. Damit wird ausgedrückt, daß ausschließlich diejenigen konzeptuellen Relationen, auf die die Dependenzrelation *subject* abgebildet wird, bei der semantischen Interpretation berücksichtigt werden. Für ein PP-Adjunkt eines transitiven Verbs, *ppadjunct*, lassen sich folgende Beschreibungen festmachen: $D_+^{\langle \text{verbtrans}, \text{ppadjunct} \rangle} := \emptyset$ und $D_-^{\langle \text{verbtrans}, \text{ppadjunct} \rangle} := \{\text{subject}, \text{dir-object}, \text{indir-object}\}$. Da ein PP-Adjunkt keine Kasusobjekte einführen kann, besagt diese Beschreibung, daß diejenigen konzeptuellen Relationen, auf die die Dependenzrelationen *subject*, *dir-object*, *indir-object* abgebildet werden, von der weiteren semantischen Interpretation auszuschließen sind.

Auf der Ebene syntaktisch induzierter Beschränkungen für die semantische Interpretation lassen sich grundsätzlich 3 Fälle unterscheiden.

1. Das aus der Syntax ableitbare Wissen *determiniert* die semantische Interpretation dann, wenn $D_+^{\text{lexval}} \neq \emptyset$ und $D_-^{\text{lexval}} = \emptyset$ (zum Beispiel für die *subject*-Valenz eines Verbs);
2. Das aus der Syntax ableitbare Wissen *restringiert* die semantische Interpretation

dann, wenn $D_+^{lexval} = \emptyset$ und $D_-^{lexval} \neq \emptyset$ (zum Beispiel bei Präpositionalphrasen als Verbkomplement);

3. Wenn $D_+^{lexval} = \emptyset$ und $D_-^{lexval} = \emptyset$, existieren keine syntaktischen Constraints und die semantische Interpretation erfolgt ausschließlich konzeptbasiert, d.h. auf der Grundlage des vorhandenen Weltwissens (zum Beispiel bei den meisten Genitivattributen).³

Um die an syntaktische Relationen gebundenen Beschränkungen auf die konzeptuelle Ebene zu übertragen, definieren wir eine Funktion $i: \mathcal{D} \rightarrow 2^{\mathcal{R}}$, die Dependenzrelationen auf Mengen von konzeptuellen Relationen abbildet. Diese Abbildungsoperation ist in der Abb. 4.3 dargestellt (beispielsweise ist $i(\text{subject}) := \{\text{AGENT}, \text{PATIENT}\}$). Für Dependenzrelationen $\delta \in \mathcal{D}$, die nicht a priori mit konzeptuellen Relationen verbunden werden können, ergibt sich $i(\delta) := \emptyset$. Die konzeptuellen Restriktionen für R_+ und R_- sind aus D_+^{lexval} und D_-^{lexval} zu berechnen, indem die Interpretationsfunktion i auf jedes Element, das in den entsprechenden Mengen enthalten ist, angewendet wird. Dadurch ergibt sich $R_+ := \{y \mid x \in D_+^{lexval} \wedge y \in i(x)\}$ und $R_- := \{y \mid x \in D_-^{lexval} \wedge y \in i(x)\}$.

In Anschluß an die statische Interpretation erfolgt die dynamische Interpretation, bei der die aus der statischen Interpretation berechneten Constraints mit dem vorhandenen Domänenwissen verrechnet werden. Diese Berechnung besteht aus einer gerichteten Suche nach einem gültigen Pfad zwischen den involvierten Konzepten unter Berücksichtigung der aus der Syntax ableitbaren Suchrestriktionen. Die Standardinterpretationsfunktion 4.1 si gibt den hier dargestellten Zusammenhang formal wieder:

$$si : \begin{cases} \mathcal{F} \times 2^{\mathcal{R}} \times 2^{\mathcal{R}} \times \mathcal{F} \rightarrow 2^{R_{con}} \\ (C_{start}, R_+, R_-, C_{end}) \mapsto \widetilde{R_{con}} \end{cases} \quad (4.1)$$

Hierbei beschreibt si die Abbildung zweier Konzepte C_{start} und C_{end} , die den spezifischsten Konzepttyp zweier in einem minimalen interpretierbaren Subgraphen enthaltener Inhaltswörter w_{start} und w_{end} darstellen, auf eine Menge von konnektierten Relationspfaden im Konzeptgraphen. Zur Bestimmung des spezifischsten Konzepttyps einer Instanz definiere ich die Funktion $type$. Beim Einlesen eines Wortes w im Eingabesatz wird zunächst eine Instanz I_w für jeden mit w im Lexikon assoziierten Konzeptidentifizier con erzeugt, so daß gilt $type(I_w) = w.con$. D.h., der spezifischste Typ einer in der Textwissensbasis neu

³Für den vierten möglichen Fall mit $D_+^{lexval} \neq \emptyset$ und $D_-^{lexval} \neq \emptyset$ gibt es keine empirischen Evidenzen, weshalb er an dieser Stelle nicht betrachtet wird.

erzeugten Instanz ist initial mit dem Typ des vom Lexikon zugewiesenen Konzeptidentifiers identisch. Jedoch kann sich aufgrund der thematische Progression der Typ einer Instanz während der Analyse verändern (vgl. Kap. 5.1). Seien in einem minimalen interpretierbaren Subgraphen die beiden Inhaltswörter w_{start} und w_{end} enthalten. Den spezifischsten Konzepttypen erhält man nun, indem man die Funktion $type$ auf die jeweiligen Instanzidentifier $I_{w_{start}}$ und $I_{w_{end}}$ anwendet. Damit ist $C_{start} =: type(I_{w_{start}})$. Ein konnektierter Relationspfad $rel_{con} \in R_{con}$ ist dann definiert durch:

$$rel_{con}((r_1, \dots, r_n)) : \Leftrightarrow \forall i \in \{1, \dots, n-1\} : \\ isa_{\mathcal{F}}(type(range(r_i)), type(domain(r_{i+1})))$$

Ein Relationspfad wird dann als konnektiert bezeichnet, wenn für alle seine n konstituierenden Relationen der Konzepttyp der Domäne von der Relation r_{i+1} den Konzepttyp des Range der Relation R_i subsumiert. Für die Pfadsuche gelten noch eine Reihe anderer Wohlgeformtheitskriterien, wie zum Beispiel Azyklizität oder der Ausschluß von Pfadmustern, welche in bereits berechneten inkludiert sind (siehe Markert & Hahn (1997)).

Um nun eine semantische Interpretation zu berechnen, löst die Interpretationsfunktion si eine Suche im azyklischen und gerichteten Konzeptgraphen aus und ermittelt alle konnektierten Pfade von C_{start} nach C_{end} . Aufgrund möglicherweise existierender konzeptueller Ambiguitäten (vgl. Kap. 4.3.7) bei der Interpretation kann mehr als ein solcher Pfad gefunden werden. Daher erfolgt die Abbildung auf die Potenzmenge von R_{con} . Diese Suche nach konnektierten Pfaden wird jedoch durch die in $R_+ \in 2^{\mathcal{R}}$ und in $R_- \in 2^{\mathcal{R}}$ enthaltenen konzeptuellen Relationen eingeschränkt. Beide Mengen von konzeptuellen Relationen reflektieren – wie bereits beschrieben – die aus den speziellen Dependenzrelation oder aus intervenierenden lexikalischen Knoten ableitbaren Constraints bezüglich der Relationierbarkeit zweier Konzepte, indem R_+ die a priori erlaubten Relationierungen und R_- die a priori ausgeschlossenen Relationierungen beinhaltet. Folglich ist $rel \in \widetilde{R_{con}}$, wenn rel einen konnektierten Relationspfad von C_{start} to C_{end} darstellt und zugleich den aus R_+ und R_- ableitbaren Beschränkungen genügt.⁴ Weiterhin werden die gefundenen Pfade, die gleichzeitig wörtliche wie figurative Lesarten beinhalten (Hahn & Markert, 1997), einem Evaluierungsprozeß unterworfen, der auf einer präferentiell geordneten Selektionsheuristik basiert (Markert, 1999). Beispielsweise ist die wörtliche Lesart gegenüber der bildlichen zu

⁴Der Sachverhalt, wie die Wohlgeformtheitsbedingungen bei der Pfadsuche Berücksichtigung finden, wird im weiteren Verlauf dieses Kapitels noch detaillierter dargestellt.

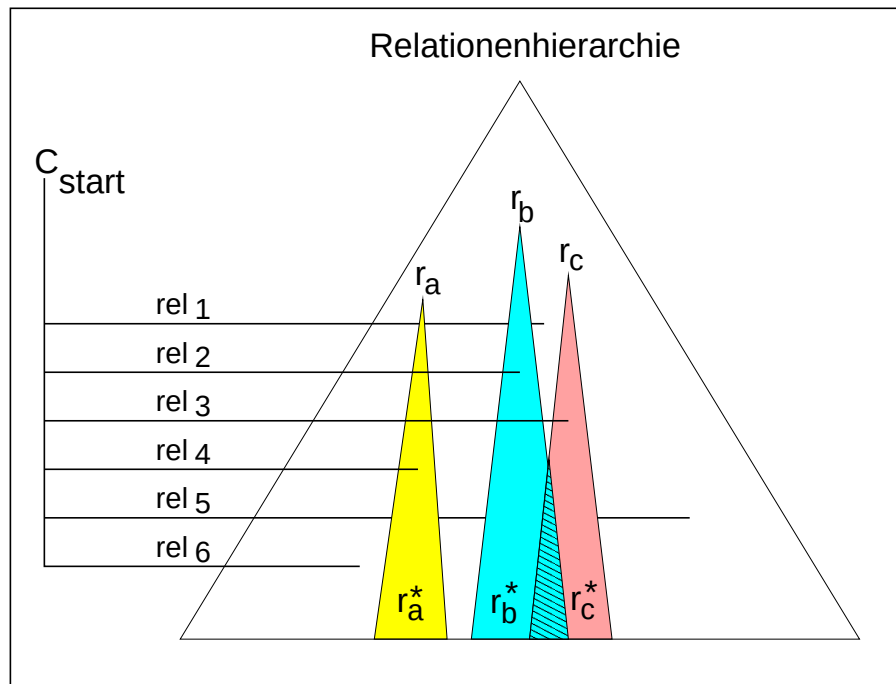


Abbildung 4.4: Interaktion der Konzeptdefinitionen mit den Restriktionslisten bei der semantischen Interpretation

bevorzugen. Darüber hinaus erfüllen einige Pfade zwar Kennektivitätskriterium, sie sind jedoch bezüglich des vorhandenen Domänenwissens implausibel. Sie werden daher den Pfaden, die bildliche Sprache repräsentieren, nachgeordnet.

Wenn die Funktion si die leere Menge zurückgibt, wenn also kein gültiger Pfad berechnet werden konnte, wird syntaxseitig keine Dependenzrelation etabliert (vgl. Kap. 3.2.4). Ansonsten ergibt jeder gültige Relationspfad $rel_i \in \widetilde{R}_{con}$ das Hinzufügen eines entsprechenden assertionalen Axioms zwischen den betroffenen Instanzen, $I_{w_{start}}$ und $I_{w_{end}}$, in der Textwissensbasis durch $(I_{w_{start}} rel_i I_{w_{end}})$, wobei rel_i die i -te Lesart denotiert (die formale Behandlung dieser interpretatorischen Ambiguitäten findet sich ausführlich in Kap. 4.3.7).

Zur Pfadsuche extrahiert die Funktion $get-roles(C) =: CR$ alle Relationen und Attribute CR von einem Konzept C in der Generalisierungshierarchie, die für dieses Konzept definiert sind. Selbstverständlich werden dabei auch solche Relationen und Attribute berücksichtigt, die von Elternkonzepten geerbt werden. Mit $get-roles(C_{start})$ lassen sich folglich die Relationen und Attribute von C_{start} ermitteln. Diese dienen dann als Startpunkte für die Suche innerhalb des Konzeptgraphen. Jedoch müssen sie zunächst noch mit

den aus der Syntax abgeleiteten Restriktionen für die Interpretation, die in R_+ und R_- enthalten sind, abgeglichen werden.

Aus Gründen der Abstraktion und Spezifikationsökonomie enthalten R_+ und R_- nur sehr allgemeine, d.h. in der Relationshierarchie sehr weit oben platzierte Relationsbezeichner. Daher sind die konkreten Rollen CR bezüglich eines Konzepts C_{start} i.a. nicht direkt mit den in R_+ und R_- enthaltenen abgleichbar. Deshalb muß vor der semantischen Interpretation die transitive Hülle für die Relationenmengen R_+ und R_- berechnet werden, um auch deren Unterrelationen mit einzuschließen. Diese berechnet sich aus $R_+^* := \{ r^* \in \mathcal{R} \mid \exists r \in R_+ : r^* \text{ isa}_{\mathcal{R}} r \}$ bzw. $R_-^* := \{ r^* \in \mathcal{R} \mid \exists r \in R_- : r^* \text{ isa}_{\mathcal{R}} r \}$. Schließlich restringiert R_+^* die Suche auf solche Relationen, die in $CR \cap R_+^*$ enthalten sind, *gdw* R_+ nicht leer ist (ansonsten sind alle Relationen aus CR zulässig). R_- hingegen erlaubt nur solche Relationen als Ausgangspunkt für die Pfadsuche, die in $CR \setminus R_-^*$ enthalten sind.

Abb. 4.4 illustriert nochmals den aufgezeigten Zusammenhang. Auf der linken Seite ist das Konzept C_{start} dargestellt. Dieser Knoten im Konzeptsystem wird durch Anwendung der Funktion *type* auf den Referenten in der Textwissensbasis ($I_{w_{start}}$) bestimmt, welcher die lexikalische Oberfläche im Satz repräsentiert. Mit Hilfe der Funktion *get-roles* lassen sich alle ererbten sowie in der Definition enthaltenen Relationen extrahieren, welche hier beispielhaft durch die Relationsbezeichner $rel_1, \dots, rel_6$ gegeben sind. Auf der rechten Seite sei die Relationshierarchie dargestellt. Jede zum Konzept C_{start} gehörige Relation hat nun einen Ort in der Relationshierarchie, der in dieser durch das Ende der Linie markiert ist. Betrachten wir nun die Restriktionsmengen R_+ und R_- . Seien rel_a, rel_b und rel_c Relationen, wie sie in diesen Listen enthalten sind. Die transitiven Hüllen dieser Relationen, rel_a^*, rel_b^* und rel_c^* sind nun die grau unterlegten Dreiecke innerhalb der Relationenhierarchie, deren oberster Punkt durch jeweils durch r_a bzw. r_b bzw. r_c markiert ist. Unterscheiden wir beispielhaft zwei mögliche Konstellationen. Seien r_a, r_b und $r_c \in R_+$, dann dienen nur die in den transitiven Hüllen enthaltenen Relationen als Startpunkte der Pfadsuche, im vorliegenden Fall also rel_2, rel_3 und rel_4 . Wenn umgekehrt r_a, r_b und $r_c \in R_-$, so werden gerade diese Relationen von Suche ausgeschlossen. Demzufolge dienen im Beispiel als Ausgang für die Pfadsuche die Relationen rel_1, rel_5 und rel_6 . Durch die beschriebene Methode erreicht man zweierlei. Zum einen kann der Suchraum für die Pfadsuche aufgrund der Qualität syntaktischen Wissens weiter eingeschränkt werden, zum anderen kann auch die "Bedeutung" von intervenierenden Knoten wie Präpositionen in den Interpretationsprozeß integriert werden.

Das in diesem Kapitel eingeführte abstrakte Interpretationsschema, welches stets zwei Knoten im Konzeptsystem als Start- und Endpunkt für eine Pfadsuche verwendet und dabei Restriktionen zur Begrenzung des Suchraums mitberechnet, dient als Grundlage für eine Reihe allgemeingültiger Spezialisierungen, die jeweils eine begrenzte Anzahl typischer Konstellationen minimaler interpretierbarer Subgraphen abdecken. Im folgenden werden diese nun der Reihe nach in den Unterkapiteln über *direct* und *mediated linkage* abgehandelt.

4.1.1.1 *Direct Linkage*

Konstellationen im Dependenzgraphen, die der Definition einer *direct linkage* entsprechen, zeichnen sich dadurch aus, daß ausschließlich die Restriktionen berücksichtigt werden müssen, die an die involvierte Dependenzrelation selbst geknüpft sind. Für die semantische Interpretation einer direkten Relationierung zweier Inhaltswörter gibt es ein spezialisiertes Interpretationsschema, si_{dir} , das die allgemeinen Fälle abdeckt, wobei gegenüber si C_{start} durch C_{head} und C_{end} durch C_{mod} ersetzt wird.

$$si_{dir} : \left\{ \begin{array}{l} \mathcal{F} \times 2^{\mathcal{R}} \times 2^{\mathcal{R}} \times \mathcal{F} \rightarrow 2^{R_{con}} \\ (C_{head}, R_+, R_-, C_{mod}) \mapsto \widetilde{R_{con}} \end{array} \right. \quad (4.2)$$

Bei der Bestimmung gültiger Relationierungen läuft die Suche innerhalb des Konzeptgraphen vom Konzepttypen des syntaktischen Kopfes C_{head} zum Konzepttypen des Modifikators C_{mod} . Die Constraints sind dabei wie oben bereits skizziert aus den an den Valenzen hängenden Beschreibungen abzuleiten, und das Schema ist anschließend mit den aktuellen Parametern des minimalen semantisch interpretierbaren Subgraphen zu instantiieren.

Zu Illustrationszwecken greifen wir auf den ersten interpretierbaren Subgraphen aus dem Beispielsatz (6) zurück, der aus den Inhaltswörtern “*Compaq*” und “*verkauft*” sowie der Dependenzrelation *subject* besteht. Die für die statische semantische Interpretation notwendigen und an die Dependenzrelation geknüpften Restriktionen sind aus den Beschreibungen für D_+^{lexval} und D_-^{lexval} der Wortklasse VERBTRANS ableitbar (siehe oben). Durch $i(subject)$ ergeben sich schließlich für R_+ die zulässigen konzeptuellen Relationen AGENT und PATIENT; R_- hingegen besteht aus der leeren Menge. Damit führt die Instantiierung von si_{dir} aus 4.2 zu folgendem Ausdruck:

$$si_{dir} := (SELL, \{AGENT, PATIENT\}, \emptyset, COMPAQ).$$

Im Nachfolgenden werden nun mit der Funktion *get-roles* die Rollen des Konzepts SELL extrahiert, i.e. die Relationen SELL-AGENT, SELL-PATIENT, BENEFICIENT, ... (vgl. Abb. 4.3). Da für die weitergehende Interpretation indessen nur die Relationen des Konzepts SELL zugelassen werden, die in der transitiven Hülle der Relationen AGENT und PATIENT liegen, reduziert sich der Suchraum auf die Relationen SELL-AGENT, SELL-PATIENT. Die Pfadsuche terminiert nach Typprüfungen im Wertebereich der beiden Relationen mit der einelementigen Menge {SELL-AGENT}, da Firmen als Handelnde einer Verkaufsaktion zugelassen sind. Folglich wird die syntaktische Anbindung zugelassen und in der Textwissensbasis die Assertion SELL.2 SELL-AGENT COMPAQ.1 vorgenommen (siehe Abb. 4.2). Die Interpretation des nachfolgenden direkten Objekts “Computer” erfolgt analog und resultiert in der Assertion SELL.2 SELL-PATIENT COMPUTER-SYSTEM.3 (vgl. jeweils die Ziffer 2 in Abb. 4.1 und Abb. 4.2).

Um zu demonstrieren, wie eine rein konzeptgetriebene Interpretation abläuft, was der Fall ist, wenn eine Dependenzrelation keine Restriktionen mit sich bringt, betrachten wir nun das Genitivattribut “Pentium-Klasse”, welches als Modifikator mit seinem syntaktischen Kopf “Computer” über die Dependenzrelation *genatt* relationiert ist. Da “Pentium-Klasse” zur Wortklasse der Nomen gehört, erbt es deren Restriktionseigenschaften bezüglich der statischen semantischen Interpretation, nämlich $D_+^{\langle noun, genatt \rangle} := \emptyset$ und $D_-^{\langle noun, genatt \rangle} := \emptyset$. Daraus ergibt sich, daß die Berechnung der möglichen konzeptuellen Relationierungen zwischen den Konzepttypen der betroffenen Instanzen in der Textwissensbasis allein auf der Grundlage des verfügbaren Domänenwissens erfolgt. Die Instantiierung des Interpretationsschemas für direkt verbundene Wörter im Dependenzgraph spiegelt diesen Zusammenhang wider:

$$si_{dir} := (\text{COMPUTER-SYSTEM}, \emptyset, \emptyset, \text{PENTIUM-CLASS})$$

Entsprechend dem Aktionsschema bei der dynamischen semantischen Interpretation sind nun mit *get-roles*(COMPUTER-SYSTEM) die konzeptuellen Relationen des Ausgangskonzepts für die Pfadsuche zu extrahieren. Da das in ITSYNDIKATE vorhandene Wissen auf die Informationstechnologie fokussiert ist, gibt der Funktionsaufruf eine große Anzahl an Relationen zurück (vgl. Kap. 5.2.1). Hierzu gehören zum einen für das Konzept COMPUTER-SYSTEM selbst definatorische Relationen, wie HAS-MOTHERBOARD, HAS-HARD-DISK und HAS-CASE. Es kommen zum anderen auch solche Relationen hinzu, die von Elternkonzepten geerbt werden, wie beispielsweise die von PRODUCT geerbten Rela-

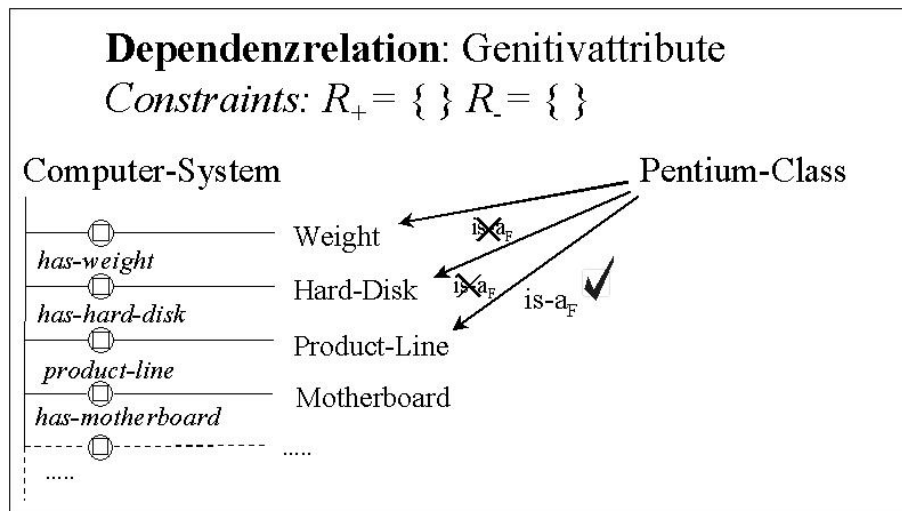


Abbildung 4.5: Semantische Interpretation eines Genitivattributs

tionen HAS-PRICE und PRODUCT-LINE oder die von PHYSICAL-OBJECT geerbten Relationen HAS-WEIGHT und HAS-PHYSICAL-DIMENSION. Für jede diese Relationen muß nun geprüft werden, ob ein gültiger Relationspfad zum Zielkonzept PENTIUM-CLASS existiert. Anschließend ermittelt der Pfadevaluierungsprozeß die plausibelste Variante. Wegen der feinkörnigen Modellierung und der enggefaßten Restriktionen für Wertebereiche an konzeptdefinitiven Relationen bleibt als Ergebnis nur eine Möglichkeit übrig, i.e. die Relation PRODUCT-LINE.⁵ Wieder wird in der Textwissensbasis die zu dieser Interpretation gehörende Assertion COMPUTER-SYSTEM.3 PRODUCT-LINE PENTIUM-CLASS.4 vorgenommen, wodurch sich für den bisher analysierten Eingabesatz die in Abb. 4.2 durch die Box 3 eingegrenzte Repräsentation ergibt.

Abb. 4.5 faßt nochmals vereinfachend zusammen, wie ein Genitivattribut semantisch interpretiert wird. In der Abbildung oben befinden sich die Restriktionslisten R_+ und R_- , welche wegen der Dependenzrelation *genatt* leer sind. Ein Teil der zuvor angeführten Relationen des Konzepts COMPUTER-SYSTEM kommt zusammen mit dem Konzepttypen ihres Wertebereichs zur Darstellung. Da für die Relation PRODUCT-LINE der Konzepttyp des Wertebereichs, PRODUCT-LINE, das Zielkonzept der Suche, PENTIUM-CLASS, subsumiert,

⁵Dieses Beispiel veranschaulicht auch, wie essentiell es ist, eine elaborierte Relationenhierarchie mit eng gefaßten Beschränkungen für den Konzepttyp an Relationen, die in Konzeptdefinitionen Verwendung finden, zu definieren. Ansonsten würden sehr viele falsch positive Ergebnisse berechnet.

kann dieses Suchergebnis schnell berechnet werden. Natürlich terminiert der Pfadsuchalgorithmus nicht, sobald die erste Lösung PRODUCT-LINE berechnet ist, dies ist in der Abbildung allerdings nicht mehr dargestellt.

Die eingeführte Interpretationsregel für *direct linkage*-Konstellationen deckt Kasus-Objekte am Verb, Genitiv-Attribute, Adverbien sowie die meisten Adjektive ab. In der Wortklasse der Adjektive gibt es allerdings noch eine Unterklasse, die eine eigene Regel erfordert, da sie von dem hier skizzierten Grundschemata für *direct linkage* abweicht. Modifizieren deverbalisierte Adjektive ein Nomen, dann füllt eigentlich die Instanz für den nominalen Kopf eine auf AGENT oder PATIENT beschränkte thematische Rolle an der Verbinstantz für den adjektivischen Modifikator.

7. Das Partikel zeigt ein vermehrt von Lymphozyten infiltriertes Stroma.

Der Satz 7 verdeutlicht den Sachverhalt. Das Substantiv “*Stroma*” wird vom Adjektiv “*infiltriertes*” modifiziert. Auf der ontologischen Ebene sind die Wortoberflächen mit dem anatomischen Konzept STROMA bzw. mit dem Verbkonzept INFILTRATION assoziiert. Das im Satz ausgedrückte Ereignis beschreibt das Einwandern von Lymphozyten in das Stroma. Entsprechend sind die LYMPHOCYTES der INFILTRATION-AGENT und das STROMA der INFILTRATION-PATIENT des Verbkonzepts INFILTRATION. Das STROMA ist also ein Slotfüller am Konzept INFILTRATION – die syntaktische Abhängigkeit verhält sich indessen genau andersherum.

$$si_{invdir} : \left\{ \begin{array}{l} \mathcal{F} \times 2^{\mathcal{R}} \times 2^{\mathcal{R}} \times \mathcal{F} \rightarrow 2^{R_{con}} \\ (C_{adjdeverbal}, R_+, R_-, C_{noun}) \mapsto \widetilde{R_{con}} \end{array} \right. \quad (4.3)$$

Das heißt, daß im Vergleich zur Standardinterpretation die Argumentreihenfolge zur Interpretation invertiert werden muß. Daher führe ich eine besondere Interpretationsfunktion si_{invdir} ein, die einen gültigen konzeptuellen Pfad zwischen dem Konzepttyp für das deverbalisierte Adjektiv und dem dazugehörigen Nomen im minimalen Subgraphen berechnet.

4.1.1.2 Mediated Linkage

Die Abdeckung an Phänomenen, die in den Bereich der *mediated linkage* fallen, erfaßt Modal- und Auxiliarkonstruktionen, Präpositionalphrasen sowie satzinterne anaphorische

Phänomene mit Relativpronomina und Possessivpronomina. Genauso wie bei der direkten Verbindung zweier Inhaltswörter über nur eine Dependenzrelation gibt es in manchen Fällen Restriktionen, die aus der Dependenzrelation abgeleitet werden können. Darüber hinaus können die auf dem Pfad liegenden intervenierenden Knoten im Dependenzgraphen Beschränkungen bezüglich der semantischen Interpretation einführen. Sollte dies – wie z.B. bei Präpositionen – der Fall sein, dann wird ein zusätzlicher Berechnungsschritt erforderlich, der die sich aus den Dependenzrelationen ableitbaren Restriktionen mit denen der intervenierenden Knoten abgleicht.

Abb. 4.1 enthält in dem durch die Box 4 eingegrenzten Teilgraphen eine Präpositionalphrase, welche sich an ihren linken Kontext anbinden möchte.⁶ Der aktuell zu interpretierende Teilgraph besteht aus den lexikalische Knoten “*Computer*”, “*mit*” und “*Festplatten*”, die durch den Kantenzug *ppattribute* und *pobject* verbunden sind. Aus Abb. 4.3 läßt sich entnehmen, daß aus der Dependenzrelation *ppattribute* keine statischen Constraints abzuleiten sind, da $D_+^{(noun, ppatt)} := \emptyset$ und $D_-^{(noun, ppatt)} := \emptyset$. Dennoch existieren statische Constraints, welche mit der Präposition “*mit*” assoziiert sind. Wir betrachten Präpositionen als Relatoren, die eine ihre “Bedeutung” kodierende Teilmenge $R_{prep} \in 2^{\mathcal{R}}$ an konzeptuellen Relationen zwischen ihrem syntaktischen Kopf und ihrem Modifikator zulassen. Folglich steht jede Präposition für eine Menge typischer Relationen zwischen den beiden Diskursentitäten im minimalen Subgraphen.

Im Gegensatz zu den an Dependenzrelationen geknüpften statischen Constraints lassen sich die an intervenierenden Knoten hängenden Beschränkungen nicht auf der Ebene der Wortklassen beschreiben, da sie die Eigenheiten eines spezifischen Worts im Lexikon kodieren. Folglich hängen diese Spezifikationen an den Blättern in der Worthierarchie, d.h. an den Wörtern selbst. Weil hiervon jedoch nur Wortbestände aus den geschlossenen Wortklassen betroffen sind, ist der Spezifikationsaufwand überschaubar.

Um zu verdeutlichen, auf welcher Ebene die Constraints im Lexikon kodiert sind, enthält Abb. 4.6 die gegenüber der Abb. 3.3 notwendigen Erweiterungen in der Beschreibung der

⁶An dieser Stelle sollte erwähnt werden, daß sich der Wortaktor für eine Präposition erst dann seinen syntaktischen Kopf sucht, wenn er seinen Modifikator gebunden hat. Damit wird der minimale interpretierbare Subgraph vollständig, sobald die Präposition ihren syntaktischen Kopf gefunden hat. Diese leichte Modifikation des inkrementellen Parsingansatzes, der die Anbindung am linken Satzkontext sofort versucht, ist deshalb sinnvoll, weil dadurch im Augenblick der Bindung der Präposition an ihren syntaktischen Kopf die semantischen Integritätsprüfungen vorgenommen werden können.

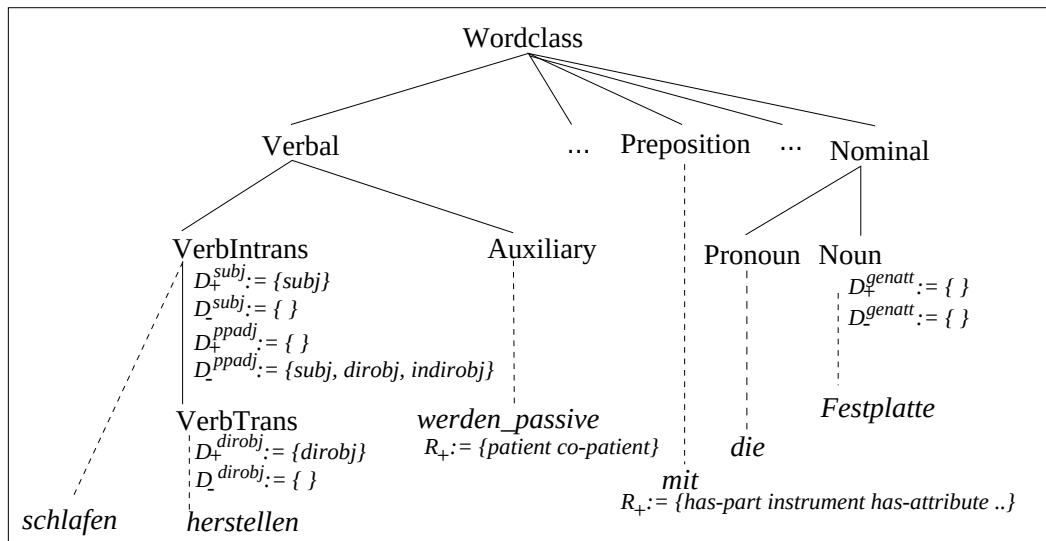


Abbildung 4.6: Erweiterung der Wortklassenhierarchie um Restriktionsbeschreibungen für die semantische Interpretation an Dependenzrelationen und intervenierenden Wörtern

Wortklassen und der lexikalischen Einträge. An Dependenzrelationen geknüpfte Spezifikationen sind mit den Wortklassen assoziiert und werden von den zur Wortklasse gehörigen Wörtern geerbt. Zum Beispiel erfolgt die Beschreibung der Beschränkungen für die Subjektsrelation, $D_+^{subject}$ und $D_-^{subject}$, an der Wortklasse VERBINTRANS (intransitive Verben) und steht durch den Vererbungsmechanismus sowohl der darunter liegenden Wortklasse VERBTRANS (transitive Verben) als auch für die eigentlichen Worteinträge an den Blättern der Wortklassenhierarchie zur Verfügung (vgl. Abb. 4.6). Weiterhin definiert die Wortklasse VERBTRANS für das vom Verb abhängige direkte Objekt $D_+^{direct-object}$ und $D_-^{direct-object}$.

Mit den intervenierenden Knoten verbundene Restriktionen sind hingegen an die Wörter selbst geknüpft, die im Wortklassenbaum die Blätter bilden und durch gestrichelte Linien angefügt sind. An der Präposition „mit“ sind die restringierenden Relationen R_+ am Worteintrag in Form von in der Relationshierarchie weit oben platzierten und damit abstrakten Relationenbezeichnern aufgelistet (siehe Abb. 4.6). Anstatt also beispielsweise sämtliche Subrelationen von HAS-PART zu enumerieren (HAS-PHYSICAL-PART, HAS-MEMBER, HAS-PHASE, HAS-CPU ...), findet sich nur die allgemeinste Relation eines interpretationsrelevanten Ausschnitt aus der Relationshierarchie in der Restriktionsmenge R_+ , nämlich HAS-PART selbst. Genau diese für lexikalische Einträgen von Wortoberflächen spezifizierten Restriktionen werden bei der Interpretation von *mediated linkage*-Subgraphen verwendet.

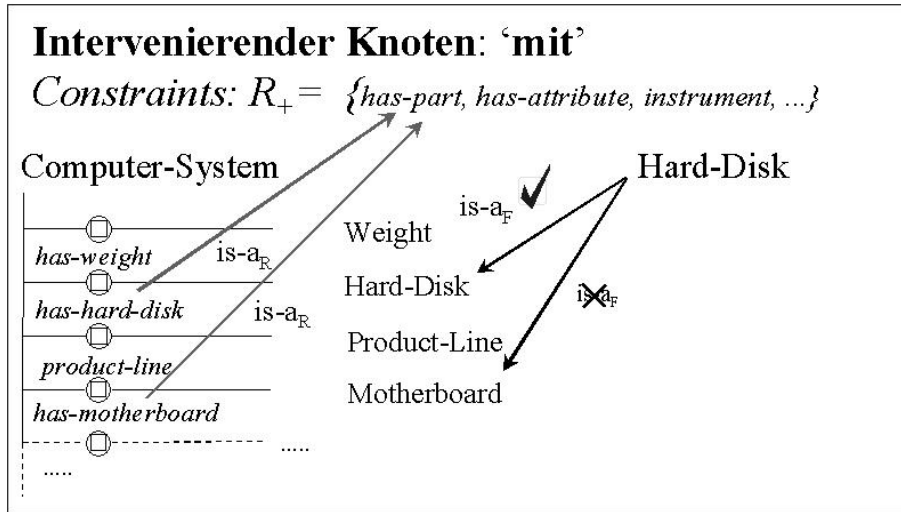


Abbildung 4.7: Semantische Interpretation einer Präpositionalphrase

Für die Interpretation von Präpositionalphrasen gibt es eine weitere Spezialisierung si_{prep} des allgemeinen Interpretationsschemas. Die Konzepttypen, für die eine Menge von gültigen Relationspfaden zu errechnen ist, sind der Konzepttyp des syntaktischen Kopfs ($C_{prephead}$) der Präposition und der Konzepttyp des von der Präposition abhängenden Modifikators (C_{prepmo}).

$$si_{prep} : \left\{ \begin{array}{l} \mathcal{F} \times 2^{\mathcal{R}} \times 2^{\mathcal{R}} \times \mathcal{F} \rightarrow 2^{R_{con}} \\ (C_{prephead}, R_+, \emptyset, C_{prepmo}) \mapsto \widetilde{R_{con}} \end{array} \right. \quad (4.4)$$

Angewandt auf den zu interpretierenden Teilgraphen im Beispielsatz (6) “Computer” – *pattribute* – “mit” – *pobject* – “Festplatten” ergibt sich folgende Instantiierung des Interpretationsschemas:

$$si_{prep} := (\text{COMPUTER-SYSTEM}, \{\text{HAS-PART}, \text{INSTRUMENT}, \dots\}, \emptyset, \text{HARD-DISK}).$$

Entsprechend der Expansion der Relationen in der Positiv-Liste R_+ auf ihre transitive Hülle sind also diejenigen Relationen des Konzepts COMPUTER-SYSTEM zulässige Ausgangspunkte für die Pfadsuche, die von $\{\text{HAS-PART}, \text{INSTRUMENT}, \text{HAS-ATTRIBUTE}, \dots\}$ subsumiert werden. Zur Verdeutlichung dieses Zusammenhangs dient Abb. 4.7. Wie bereits zuvor bei der Interpretation des Genitivattributs extrahiert die Funktion *get-roles* sämtliche Konzeptrollen des Konzepts COMPUTER-SYSTEM, jedoch wird nun ein Pfad nicht mehr für alle diese Relationen gesucht, sondern nur noch für die, die in der transitiven

Hülle von R_+ enthalten sind (hier HAS-MOTHERBOARD und HAS-HARD-DISK). Die mit $isa_{\mathcal{R}}$ gelabelten gerichteten Kanten zeigen eine Subsumptionsbeziehung zwischen den in der Grafik enthaltenen Relationen an. Weil die Relation HAS-HARD-DISK von HAS-PART subsumiert wird, stellt die Relationierung von COMPUTER-SYSTEM mit HARD-DISK eine gültige semantische Interpretation dar. In Abb. 4.2 entsteht durch Hinzufügen der Assertion COMPUTER-SYSTEM.3 HAS-HARD-DISK HARD-DISK.5 der durch die Box 4 umrissene Konzeptgraph als Inhaltsrepräsentation des entsprechenden Dependenzgraphen in Box 4 von Abb. 4.1.

In einigen Fällen erfolgt die Anbindung der Präpositionalphrase zusätzlich unter den aus einer Dependenzrelation ableitbaren Restriktionen, die vom syntaktischen Kopf der Präposition abhängt. Dies ist zum Beispiel der Fall bei einer mit „von“ eingeleiteten Präpositionalphrase, die den Agenten eines Passivsatzes einführt. Im Satz (6) betrifft dies den minimalen interpretierbaren Subgraphen, der sich aus den Wörtern „hergestellt“, „von“ und „IBM“ konstituiert. Die in unserem Satzkontext gegebene passive Lesart des Wortes „hergestellt“ erbt über die Wortklassenhierarchie die für seine Wortklasse VERBPART-PASS definierte Dependenzrelation $ppsubject$, i.e. der Subjektsfall für ein Verb im Partizip Passiv. Entsprechend finden sich bei der zum Partizip Passiv gehörenden Dependenzrelation die Restriktionen $D_+^{(verbpартpass, ppsubject)} := \{\text{SUBJECT}\}$ und $D_-^{(verbpартpass, ppsubject)} := \emptyset$. Gehen aus den Beschreibungen des syntaktischen Kopfs einer PP zusätzlich zu den durch die Präposition selbst eingeführten Restriktionen weitere Beschränkungen hervor, so müssen diese in einem zusätzlichen Berechnungsschritt miteinander abgeglichen werden. Hierzu wird durch die Anwendung der Funktion i auf alle Elemente in D_+^{lexval} bzw. D_-^{lexval} eine neue Menge $P_+ := \{y \mid x \in D_+^{lexval} \wedge y \in i(x)\}$ berechnet, die alle durch die Syntax sanktionierten konzeptuellen Relationen beinhaltet. Analog errechnet sich $P_- := \{y \mid x \in D_-^{lexval} \wedge y \in i(x)\}$. Darüber hinaus erfolgt die Expansion der Mengen P_+ und P_- auf ihre transitiven Hüllen, P_+^* und P_-^* , durch $P_+^* := \{p^* \in \mathcal{R} \mid \exists p \in P_+ : p^* isa_{\mathcal{R}} p\}$ und $P_-^* := \{p^* \in \mathcal{R} \mid \exists p \in P_- : p^* isa_{\mathcal{R}} p\}$.

Der bei der semantischen Interpretation von Präpositionen hinzukommende Berechnungsschritt prüft die Kompatibilität der statischen Constraints aus der Dependenzrelation (die durch die Anwendung von i auf D_+^{lexval} und D_-^{lexval} gegeben sind) mit den Constraints, die aus der jeweiligen Präposition als intervenierendem Knoten ableitbar sind. Bei nicht leerer Menge P_+ ist eine weitergehende Interpretation dann obsolet, wenn $(R_{Prep}^* \setminus P_-^*) \cap P_+^* = \emptyset$, weil die Verrechnung der Beschränkungen aus der Dependenzrelation des Kopfs

der Präposition mit den an der Präposition selbst spezifizierten Beschränkungen keine Interpretation zuläßt, weil die Menge der erlaubten Relationen leer ist. Ansonsten wird R_+ mit $R_+ := (R_{P_{rep}}^* \setminus P_-^*) \cap P_+^*$ ermittelt, falls $P_+ \neq \emptyset$. Falls $P_+ = \emptyset$ errechnet sich R_+ aus $R_+ := (R_{P_{rep}}^* \setminus P_-^*)$.

Im Beispiel bestehend aus dem Teildependenzgraphen *“hergestellt” – ppsubject – von* – *pobject – “IBM”* wird genau dieser zusätzliche Berechnungsschritt notwendig, da an der *ppsubject*-Dependenz-Relation des Partizips Passiv statische Interpretationsconstraints hängen. Die Dependenzrelation *subject* läßt sich statisch auf $\{\text{AGENT}, \text{PATIENT}\}$ abbilden. Die Relation *AGENT* ist im Gegensatz zu *PATIENT* auch in R_{von} enthalten, damit sind sowohl die Relation *AGENT* und alle ihre Unterrelationen in R_+ der Präposition *“von”* enthalten (i.e. der Schnittmenge aus P_+^* und $R_{P_{rep}}^*$). Da wegen $R_+ \neq \emptyset$ eine dynamische semantische Interpretation getriggert werden kann, wird das semantische Interpretationsschema für Präpositionen folgendermaßen instantiiert:

$$si_{prep} := (\text{MANUFACTURE}, \{\text{AGENT}\}, \emptyset, \text{IBM}).$$

Im nachfolgenden Interpretationsschritt werden nur solche Rollen betrachtet, die in der Schnittmenge von $get\text{-}roles(\text{MANUFACTURE})$ und R_+ liegen, was sich, gegeben die Ontologie, auf die Relation *MANUFACTURE-AGENT* isa_R *AGENT* eingeschränkt. Daher reduziert sich die Erfüllung der dynamischen Constraints auf die Prüfung, ob unter der Relation *MANUFACTURE-AGENT* ein gültiger Pfad zu *IBM* berechnet werden kann. Da *IBM* von der Typrestriktion *PERSON* der Relation *MANUFACTURE-AGENT* über die Zwischenkonzepte *LEGAL-PERSON* und *COMPANY* subsumiert wird, läßt sich die Assertion *MANUFACTURE.7 MANUFACTURE-AGENT IBM.6* hinzufügen (vgl. Abb. 4.2 Box 5).

Aus einer Dependenzrelation ableitbare syntaktische Wohlgeformtheitsbedingungen sind bei der Anbindung von Präpositionalphrasen nicht nur für passivische Konstruktionen festzumachen. Eine große Anzahl von Verben führt (oftmals obligatorische) Valenzen nicht über die Dependenzrelationen *direct-object* oder *indirect-object* ein, sondern über Präpositionalphrasen (vgl. Dahlgren & McDowell (1986), Gomez et al. (1997)). Als Beispiel sei der folgende Satz aus der Informationstechnologie angeführt:

8. Mit Hochdruck arbeiten die Hersteller an besserer Performance.

Das Objekt, an dem *“die Hersteller”* arbeiten (die *“Performance”*), wird nicht durch einen Kasus, sondern durch eine Präpositionalphrase eingeführt. Da Präpositionalphrasen dieser Klasse entweder eine obligatorische Valenz am Verb einführen oder zumindest,

wenn sie nicht obligatorisch sind, eine eindeutig bestimmbare Beschränkung bezüglich der semantischen Interpretation mit sich bringen, muß dieses Restriktionswissen dem Interpretationsprozeß zugeführt werden. Zum einen sind hierfür die Restriktionslisten R_+ an den jeweiligen Präpositionen in der Wortklassenhierarchie durch Aufnahme der thematischen Rollen entsprechend zu adaptieren, d.h., PATIENT gehört folglich zur Restriktionsmenge von „an“. Zum anderen sind die Wörter selbst in der Wortklassenhierarchie um die Beschreibungen ausgezeichneter Präpositionalphrasen zu erweitern, die Kasusobjekte einführen. In der Grammatikspezifikation ist daher zur Abdeckung dieses Phänomens eine spezielle Präpositionalphrasen-Valenz an Verben, aus Verbnominalisierungen abgeleiteten Substantiven und deverbalsierten Adjektiven definiert (*ppcomp*). Mit dieser ist demnach ein als Kasusobjekt zu interpretierendes Komplement gekennzeichnet, welches über eine Präpositionalphrase an die drei bezeichneten Wortklassen angebunden wird.

Die Notwendigkeit, präpositionale Komplemente in der Wortklassenhierarchie auf der Ebene der Worteinträge, also auf der Ebene der Blätter, vornehmen zu müssen, widerspricht mitnichten dem angestrebten Ideal, ein Maximum an Abstraktion in den Beschreibungen zu erreichen. Diese Form der Präpositionalphrasen als Komplement sind ausnahmslos mit speziellen Lexikoneinträgen verbunden und weisen bezüglich ihrer Distribution weder intra- noch intersprachlich irgendwelche Regelmäßigkeiten auf, die zur Abstraktion verwendet werden könnten.⁷ Ein weiteres Merkmal des Verbkomplements ist, daß es syntaktisch – ebenso wie die *subject*, *direct-object* und die *indirect-object* Valenz – nur einmal auftreten kann (vgl. hierzu (Bröker, 1999)).

Im Gegensatz zu den Komplementen erfüllen präpositionale Adjunkte keine Funktion bezüglich des Valenzinventars eines Verbs, eines Adjektivs oder eines Nomens, d.h., ihre Verwendung ist rein fakultativ. Darüber hinaus gibt es keinerlei numerische Beschränkung für die Häufigkeit der Okkurrenz eines PP-Adjunkts, das durch die gleiche Präposition eingeführt wird. Die Grammatikspezifikation trägt diesem Sachverhalt dadurch Rechnung, daß sie neben dem in Dependenzgrammatiken gängigen Begriff der Dependenz zusätzlich das Konstrukt der *Vakanz* einführt (Bröker, 1999). Vakanzen bezeichnen Leerstellen

⁷Dies ist auch der augenscheinliche Grund dafür, daß beim Fremdsprachenerwerb neben dem Lernen von rein idiomatischen Ausdrücken Verben grundsätzlich zusammen mit den dazugehörigen Präpositionen gelernt werden sollten, da sonst die erworbene Information unvollständig ist und damit eine ständige Fehlerquelle darstellt. Bedauerlicherweise ignorieren eine Reihe von Fremdsprachwörterbüchern diesen Sachverhalt.

an einem syntaktischen Kopf, die fakultativ gefüllt werden können und zugleich mehrere gleichartige Dependenzrelationen am selben syntaktischen Kopf zulassen. Darüber hinaus substituieren diese Modifikatoren keine obligatorischen Dependenz eines Valenzrahmens (wie dies zum Beispiel die *ppsubject*-Valenz in einem Passivsatz tut). Zur Illustration mag das nachfolgende, zugegebenermaßen etwas unnatürlich wirkende Beispiel dienen:

9. Die Aufrüstung von Apple-Computern von OS1.2 auf OS2.0 vom vergangenen Monat wurde erfolgreich abgeschlossen.

Die über die Präposition „von“ eingeführten Präpositionalphrasen, welche eine adverbiale Bestimmung der Handlung „Aufrüstung“ beinhalten, nämlich „von OS1.2“ und „vom vergangenen Monat“, füllen keine thematische Rolle am Verbkonzept UPGRADE. Im Gegensatz dazu führt die Präpositionalphrase „von Apple-Computern“ das von der Aufrüstung betroffene Objekt ein. Deutlicher wird der Zusammenhang in der um die Verbnominalisierung bereinigten Paraphrase von Satz 9: „Im vergangenen Monat wurden Apple-Computer von OS1.2 auf OS2.0 aufgerüstet.“ Im paraphrasierten Satz füllt „Apple-Computer“ die *subject*-Valenz des Passivsatzes und ist damit ein Kasusobjekt. Im Ursprungssatz ist die Präpositionalphrase „von Apple-Computern“ keineswegs obligatorisch. In der Konsequenz ist Obligatheit kein Kriterium, um zu bestimmen, ob eine Komplement- oder eine Adjunktlesart vorliegt. Das in meinem Ansatz zur semantischen Interpretation verwendete Kriterium kombiniert die Wortklassenzugehörigkeit mit dem Konzeptsystem. Wenn eine Wortoberfläche ein Verb, ein deverbaliertes Adjektiv oder ein deverbaliertes Substantiv ist und eine Präpositionalphrase am dazugehörigen Verbkonzept eine konzeptuelle Relationen füllt, die sich aus den Dependenzrelationen *subject*, *direct-object* und *indir-object* ableiten lassen, dann handelt es sich bei der Präpositionalphrase um ein Komplement, ansonsten um ein Adjunkt.

Für Präpositionalphrasen als Vakanzen gibt es indessen keine determinierte Lesart. Dennoch besteht auf der Grundlage der im Vortext aufgezeigten Zusammenhänge die Möglichkeit, wenigstens umgekehrt festzulegen, welche Interpretationsvarianten auszuschließen sind. Um diese Differenzierung über einen eindeutigen syntaktischen Relationsbezeichner zu realisieren, ist hierfür die Dependenzrelation *ppadjunct* (für Präpositionaladjunkt) definiert. Die Constraints für die semantische Interpretation sind allerdings über Ausschlusskriterien gegeben, die festlegen, welche Interpretationen unzulässig sind. Entsprechend ist in der Wortklassenhierarchie für VERBINTRANS kodiert: $D_+^{(verbintrans, ppadjunct)} := \emptyset$ und

$D_-^{\langle \text{verbintrans}, \text{ppadjunct} \rangle} := \{\text{subject}, \text{direct-object}, \text{indirect-object}\}$. D.h., falls eine Präpositionalphrase als Adjunkt auftritt, dann müssen gemäß dem Interpretationsschema diejenigen konzeptuellen Relationen P_- von der dynamischen semantischen Interpretation ausgeschlossen werden, die sich durch die statische Interpretation i aus $D_-^{\langle \text{verbintrans}, \text{ppadjunct} \rangle} := \{\text{subject}, \text{direct-object}, \text{indirect-object}\}$ ermitteln lassen. Durch Komplementbildung der transitiven Hülle dieser Relationenmenge mit der transitiven Hülle der an der Präposition hängenden Relationen $r \in R_+$ berechnen sich die zulässigen Relationen zur dynamischen semantischen Interpretation. Das Komplement ergibt sich aus: $R_+ := (R_+^* \setminus P_-^*)$. Damit kann von vorneherein ausgeschlossen werden, daß die PP „von OS1.2“ im Kontext des Beispielsatzes beispielsweise als AGENT, PATIENT, CO-PATIENT interpretiert wird. Die Synthese unterschiedlicher Quellen von restringierendem Wissen, i.e. die Dependenzrelation des syntaktischen Kopfs der Präposition wie die durch die Präposition selbst eingeführten Constraints, führt zu einer Begrenzung des Suchraums bei der dynamischen semantischen Interpretation und damit auch zu einer Begrenzung ambiger Lesarten bei der Relationierung zweier Diskursobjekte in der Textwissensbasis.

Die nächste Konstellation eines minimalen semantisch interpretierbaren Subgraphen, die in Satz 6 enthalten ist, zieht zwei unterschiedliche Phänomenklassen zusammen. Zum einen beinhaltet sie die Interpretation einer passivischen Auxiliarkonstruktion, zum anderen kann mit ihrer Hilfe veranschaulicht werden, wie die semantische Interpretation mit der Auflösung anaphorischer Phänomene interagiert. Es handelt sich um den Subgraphen („die“ – subject – „werden“ – *vrbpart* – „hergestellt“). Abb. 4.1 grenzt diesen Graphen durch die Box mit der Nummer 6 ein.

Zunächst stelle ich das spezielle semantische Interpretationsschema für Auxiliarkonstruktionen vor. Dieses kommt übrigens gleichermaßen für komplexere Modalkonstruktionen wie auch für Kombinationen aus beiden zur Anwendung, da die Basiskonfiguration für alle diese Fälle die gleiche ist.⁸ Die nachfolgenden Ausführungen gehen zunächst indessen nur auf die Passivinterpretation ein. Für diese gilt: Die Instanz für das Subjekt $I_{w_{\text{subject}}}$ im Subgraphen, das über die *subject*-Dependenzrelation mit der die Tempusmarkierung tragenden Auxiliarform verbunden ist, muß mit der Instanz für das eigentliche bedeutungstragende Verb $I_{w_{\text{vrbpart}}}$ im Subgraphen durch den Interpretationsprozeß relationiert

⁸Mit komplexeren Konstruktionen sind minimale Subgraphen gemeint, die aus mehr als drei Knoten bestehen und modale wie auxiliare Verben kombinieren. Bsp.: „Der Computer hätte früher auf den Markt gebracht werden können.“.

werden. Dieses Verb erscheint als Partizipform im Partizip Passiv. Im passivierten Satz wird es über die Dependenzrelation $vrupart$ in den minimalen semantisch interpretierbaren Subgraphen eingebunden. Hierzu sind jeweils der spezifischste Konzepttyp für beide Diskursobjekte zu ermitteln, daher ergeben sich $C_{vrupart} := type(I_{w_{vrupart}})$ und $C_{subject} := type(I_{w_{subject}})$. Die bei der semantischen Interpretation zu berücksichtigenden Restriktionen $R_{passaux} := \{PATIENT, CO-PATIENT\}$ stammen vom intervenierenden Knoten, dem Passivauxiliar “werden”. Wie bei allen intervenierenden Knoten sind sie auf der Ebene des lexikalischen Eintrags spezifiziert (siehe Abb. 4.6). Es ergibt sich somit für die semantische Interpretation des Passivs das nachfolgende Interpretationsschema:

$$si_{aux} : \left\{ \begin{array}{ll} \mathcal{F} \times 2^{\mathcal{R}} \times 2^{\mathcal{R}} \times \mathcal{F} & \rightarrow 2^{R_{con}} \\ (C_{vrupart}, R_{passaux}, \emptyset, C_{subject}) & \mapsto \widetilde{R_{con}} \end{array} \right. \quad (4.5)$$

Gesucht wird im Konzeptgraphen also nach einer Relationierung zwischen dem Konzepttypen der Verbinstantz und dem Konzepttypen des syntaktischen Subjekts. Bezogen auf den Beispielsatz (6) führt dies zur Instantiierung des speziellen Interpretationsschemas mit den aktuellen Parametern im Satz:

$$si_{aux} := (\text{MANUFACTURE}, \{PATIENT, CO-PATIENT\}, \emptyset, \text{HARD-DISK}).$$

Die Pfadsuche terminiert mit der einelementigen Menge $\{\text{MANUFACTURE-PATIENT}\}$ als Ergebnis. Die Textwissenbasis wird um die dazugehörige Assertion $\text{MANUFACTURE.7 MANUFACTURE-PATIENT HARD-DISK.5}$ erweitert (siehe Box 6 in Abb. 4.2), womit schließlich die Interpretation des Eingabesatzes abgeschlossen ist.

Allerdings hat die vorangegangene Darstellung das Faktum unterschlagen, daß der Referent des im Satz verwendeten Relativpronomens “die” zu Beginn nicht bekannt ist. Die Verwendung von Proformen, mit welchen kein Konzeptidentifizier verbunden ist, führt zu einer Modifikation des Begriffs eines “minimalen interpretierbaren Subgraphen”. Nach den bisherigen Kriterien ist der Subgraph (“die” – subj – “werden” – vrupart – “hergestellt”) kein semantisch interpretierbarer Subgraph, da mit dem Relativpronomen ein Tupel an zu relationierenden Diskursobjekten, wie es grundsätzlich benötigt wird, zunächst nicht zur Verfügung steht.

Definition 4.2 *Ein minimaler semantisch interpretierbarer Subgraph nach Definition 4.1 heißt unvollständig, wenn w oder w' kein Inhaltswort, sondern eine Proform ist.*

Ein minimaler semantisch interpretierbarer Subgraph nach Definition 4.1 heißt vollständig,

wenn w und w' Inhaltswörter sind oder, falls w und/oder w' Proformen sind, die Referenten für w und/oder w' verfügbar sind.

Die Berechnungen, die die Referenten der Proform satzintern ermitteln, kommen hier nicht zur Darstellung, sondern werden vorausgesetzt (Strube & Hahn, 1995b). Die Resolution satzübergreifender, auf Proformen basierender anaphorischer Phänomene wurde in Kap. 3.2.5 behandelt.

Wenn ein minimaler interpretierbarer Subgraph unvollständig ist, dann muß die semantische Interpretation solange aufgeschoben werden, bis der jeweilige Referent verfügbar gemacht wird. Da sich bei Satz (6) der Referent des verwendeten Relativpronomens im selben Satz befindet, kann die Interpretation nicht stattfinden, bevor sich der relative Nebensatz über die Dependenzrelation *rela* an den Matrixsatz angebunden hat. Diese Dependenzrelation bindet sich an den Kopf des gesamten Nebensatzes, in unserem Fall also *“werden”*. Durch die Endständigkeit dieses Wortes kann erst am Satzende auf den Matrixsatz zur Suche nach einem Referenten zugegriffen werden. Mit der Abbildung von *“Festplatten”* auf das Relativpronomen *“die”* im Anschluß an die Prüfung einer Reihe von Kongruenzbedingungen wird das Diskursobjekt HARD-DISK.5 in den Subgraphen an die Stelle des *“Platzhalters”* Relativpronomens eingefügt. Damit liegt ein vollständiger semantisch interpretierbarer Subgraph vor, dessen Interpretationsergebnis bereits bekannt ist.

Dieses kurze Beispiel der Resolution einer satzinternen anaphorischen Referenz illustriert, wie verschiedene Wissens Ebenen bei der semantischen Interpretation zusammenspielen und dabei lokale wie globalere Berechnungsprozesse integrieren.

Beim gegenwärtigen Stand der Implementation deckt die semantische Interpretation zu *mediated linkage* alle Formen von Präpositionalphrasen ab, unabhängig davon, ob sie ein Verb, ein Adjektiv oder ein Substantiv als syntaktischen Kopf haben. Darüber hinaus sind passivische Konstruktionen ebenso erfaßt wie Konstruktionen mit Modal- oder Auxiliärverben.

4.1.2 Die semantische Ebene

Bei der Betrachtung der semantischen Ebene sind die beiden syntaktischen Konstellationen von *direct* und *mediated linkage* grundsätzlich zu unterscheiden. Im Fall einer *direkt linkage* kann die semantische Interpretation unmittelbar mit den konzeptuellen Korrelaten erfolgen, da alle für die Instantiierung der entsprechenden semantischen Interpretationsfunktion

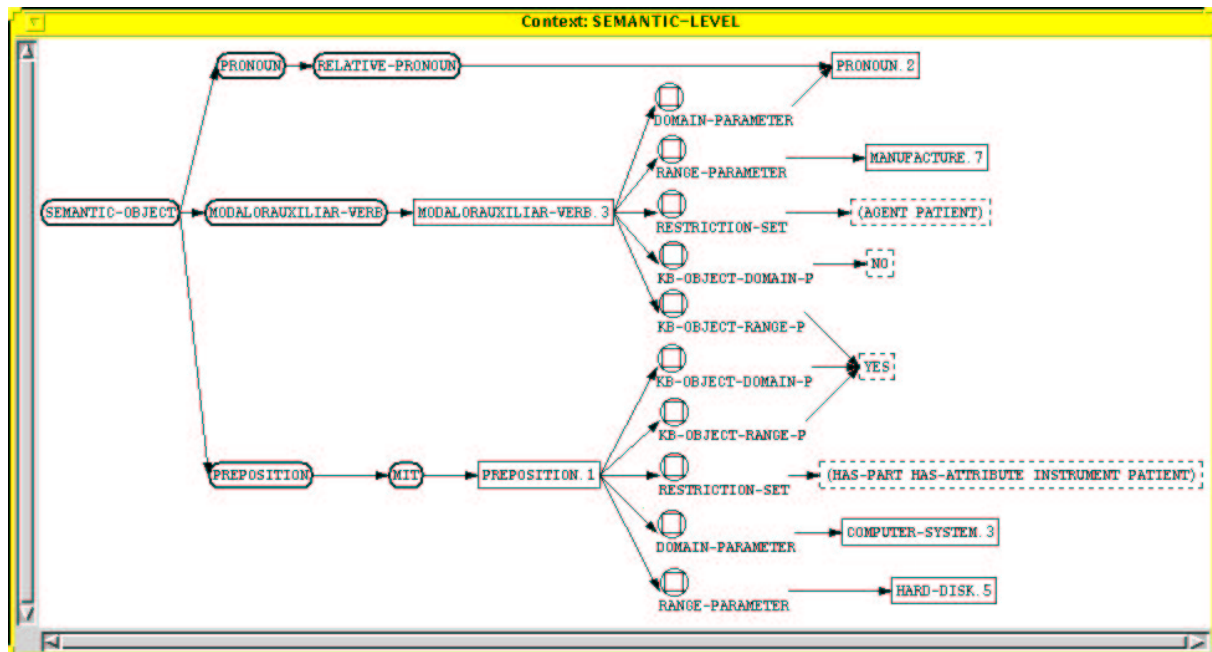


Abbildung 4.8: Semantische Interpretation mit semantischen Objekten

notwendigen Informationen vorhanden sind: die beiden beteiligten Diskursentitäten (i.e. die Instanzen in der Textwissensbasis) sowie die beiden Restriktionsmengen.

Im Fall der *mediated linkage* verhält es sich anders, da an den intervenierenden Knoten zusätzliche Constraints hängen und die dazugehörigen syntaktischen Subgraphen durch die inkrementelle Analyse nicht von Anfang an vollständig zur Verfügung stehen. Um jedoch die schon nutzbaren Teilberechnungsergebnisse zu administrieren, existiert eine mit der Textwissensbasis verbundene semantische Ebene. In dieser besonderen Ebene ist eine Anzahl semantischer Objekte definiert, die dem Zweck dienen, Teilberechnungsergebnisse zu repräsentieren, solange der minimale interpretierbare Subgraph unvollständig ist.

Die semantische Ebene verwaltet also die in minimalen Subgraphen auftretenden Weltwissenobjekte und die dazugehörigen Restriktionslisten. Die semantische Ebene ist wegen der engen Kopplung an die Textwissensbasis in demselben Formalismus implementiert wie das Upper-Model und die MED- bzw. IT-Wissensbasis, d.h. in terminologischer Logik. Intervenierende Knoten innerhalb der minimalen Subgraphen sind in Form von semantischen Objekten gegeben, die die Diskursentitäten der Textwissensbasis in einer eigenen Partiti-

on manipulieren.⁹ Abgesehen von den Restriktionslisten haben sie i.a. die gleiche Struktur. Um die Vollständigkeit der zu einem minimalen Subgraphen gehörenden Assertionen in der Textwissensbasis zu kontrollieren, ist zunächst ein Konzept SEMANTIC-OBJECT definiert. Die zur semantischen Interpretation von unterschiedlichen Sprachphänomenen notwendigen Spezialisierungen (z.B. MODALORAUILIAR-VERB oder PREPOSITION) sind Subkonzepte von SEMANTIC-OBJECT. Zum semantischen Objekt PREPOSITION gibt es noch eine zur Anzahl der Wortoberflächen äquivalente Anzahl von Spezialisierungen. In Abb. 4.8 findet sich ein Ausschnitt aus dieser Konzepthierarchie, der die Präposition MIT einschließt.

Eine Sonderstellung nehmen die auch in der Abbildung enthaltenen Pronomina (PRONOUN) ein, denn sie treten syntaktisch niemals als intervenierende Knoten auf. Diese Klasse semantischer Objekte dient als "Platzhalter" in den semantischen Strukturen, bis die Referenzauflösung erfolgt ist, die mit Hilfe einer Koreferenzrelation an der Instanz angezeigt wird. Bei der semantischen Interpretation eines minimalen Subgraphen, in dem ein Pronomen als Modifikator auftritt, fügt die semantische Ebene dann den Referenten des Pronomens in Form einer konzeptuellen Relationierung zwischen den beiden involvierten Diskursentitäten in die Textwissensbasis ein.

Semantische Objekte, deren Wortoberflächen syntaktisch als intervenierende Knoten auftreten, haben jeweils eine identische Struktur. Zunächst verfügen sie über zwei Relationen, die die beiden Diskursentitäten (Instanzen) des minimalen Subgraphen aufnehmen (DOMAIN-PARAMETER und RANGE-PARAMETER) und die dann als Parameter der jeweiligen Interpretationsfunktion übergeben werden. Dann gibt es einen speziellen Slot, der die Restriktionsmengen bei der Interpretation aufnimmt (RESTRICTION-SET). Diese Relation ist unabhängig vom textuellen Kontext durch die in der Wortklassenhierarchie vorkodierten jeweiligen Mengen R_+ vorbelegt. Schließlich existieren noch die beiden Relationen KB-OBJECT-DOMAIN-P und KB-OBJECT-RANGE-P, deren Füller mit den Boole'schen Werten 'YES oder 'NO zu belegen ist. Dieser Wert gibt das Ergebnis eines Prüfprädikats an, das Auskunft darüber gibt, ob die Füller der Relationen DOMAIN-PARAMETER oder RANGE-PARAMETER Diskursentitäten der Textwissensbasis ('YES) oder aber Proformen ohne resolvierten Referenten darstellen ('NO).

Um die Funktionsweise der semantischen Ebene zu verdeutlichen, enthält die Abb.4.8 neben einem Fragment der Konzepthierarchie die Repräsentation eines Teils der semanti-

⁹Gründe, die dafür sprechen, die Repräsentationsebene für die Semantik und das Weltwissen in einem uniformen Formalismus zu halten, finden sich bei Allen (1993).

schen Objekte für den Satz 6, nämlich die für den Teilsatz: *“Computer mit Festplatten, die hergestellt werden”*. Die drei zur semantischen Interpretation notwendigen semantischen Objekte sind PREPOSITION.1 für die Präposition *“mit”*, MODALORAUXILIAR-VERB.3 für das Passivauxiliar *“werden”* und schließlich PRONOUN.2 für das Relativpronomen *“die”*. Anhand der Slotbelegungen am semantischen Objekt PREPOSITION.1 ist ersichtlich, daß der minimale interpretierbare Subgraph *“Computer mit Festplatten”* vollständig ist. Beide Instanzen in der Wissensbasis (COMPUTER-SYSTEM.3 und HARD-DISK.5) sind verfügbar. Folglich ist auch die semantische Interpretation in der Textwissensbasis bereits erfolgt. Beim semantischen Objekt des Passivauxiliars (MODALORAUXILIAR-VERB.3) verhält es sich hingegen anders. Zwar verweist der Füller für RANGE-PARAMETER auf eine Instanz in der Textwissensbasis (MANUFACTURE.7), der Füller an der Relation DOMAIN-PARAMETER hält indessen ein nicht resolviertes Pronomen (PRONOUN.2). Daß das Pronomen noch nicht resolviert ist, erkennt man an der Belegung des Slots KB-OBJECT-DOMAIN-P, die 'NO ist. Aus den um Referenzresolution erweiterten Bedingungen für einen minimalen interpretierbaren Subgraphen kann abgeleitet werden, daß die Struktur noch nicht interpretierbar ist. Bezogen auf den Verlauf der inkrementellen Analyse spiegelt die semantischen Ebene den Zustand wider, daß die Anbindung des Relativsatzes an den syntaktischen Kopf *“Festplatten”*, die zur Suche nach dem Antezedenten führt, noch nicht erfolgt ist. Erst die Auflösung des satzinternen anaphorischen Bezugs würde die Struktur vervollständigen, indem HARD-DISK.5 den Slot DOMAIN-PARAMETER belegt und entsprechend der Wert von KB-OBJECT-DOMAIN-P auf 'YES gesetzt wird.

Der Vorteil des gewählten Ansatzes, keine intermediäre Repräsentationsschicht zur semantischen Interpretation einzuziehen, liegt – abgesehen von der Verschlinkung in der Systemarchitektur – darin, daß die in der Textwissensbasis erzielten Interpretationsergebnisse der semantischen Ebene unmittelbar zur Verfügung stehen. Darüber hinaus werden vollständige interpretierbare Strukturen sofort berechnet, so daß die Textwissensbasis sich prinzipiell parallel zu den syntaktisch bereits analysierten Strukturen entwickelt.

4.1.3 Konzeptuelle Interpretation

Mit der in Definition 4.1 festgelegten Beschreibung eines minimalen interpretierbaren Subgraphen kann nunmehr die rein semantische Interpretation von der *konzeptuellen* Interpretation unterschieden werden. Während die semantische Interpretation die unmittelbare

Transformation von der syntaktischen Ebene auf die konzeptuelle Repräsentationsebene umschreibt und dabei Paare von Diskursobjekten relationiert, greift die konzeptuelle Interpretation auf die Ergebnisse dieses Interpretationsprozesses in der Textwissensbasis zurück. Bei der konzeptuellen Interpretation können im Gegensatz zur semantischen Interpretation durchaus mehr als zwei Diskursobjekte involviert sein. Da die konzeptuelle Interpretation allein auf logischen Bezügen zwischen den Assertionen fußt, die durch die semantische Interpretation in die Textwissensbasis eingetragen wurden, arbeitet sie regelbasiert. Eine Einbeziehung der Syntax erfolgt nur insofern, als die in der Textwissensbasis vorhandenen Assertionen durch die semantische Interpretation unter Integration syntaktischen Wissens in diesen Interpretationsprozeß gewonnen wurden. Bezogen auf die Definition eines minimalen interpretierbaren Subgraphen ist die konzeptuelle Interpretation gerade dadurch umschrieben, daß es auf dem Pfad zwischen zwei Inhaltswörtern w_1 und w_n , deren konzeptuelle Korrelate in der Textwissensbasis durch die Interpretation manipuliert werden, mindestens noch ein weiteres Inhaltswort gibt.

Definition 4.3 *Seien w_1 und w_n zwei Inhaltswörter in einem Satz S , die durch einen Pfad im Dependenzgraphen miteinander relationiert sind. Eine konzeptuelle Interpretation liegt genau dann vor, wenn folgende zwei Bedingungen erfüllt sind*

1. *es gibt mindestens einen Index $i \in \{1, \dots, n\}$, so daß w_i ein Inhaltswort ist.*
2. *w_1 und w_n werden durch Hinzufügung mindestens einer Assertion in der Textwissensbasis miteinander konzeptuell relationiert.*

Für die konzeptuelle Interpretation sind grundsätzlich zwei Fälle zu differenzieren. Erstens können weitere Verträglichkeitsprüfungen zwischen Füllern von Instanzrollen vorgenommen werden, die mit derselben Instanz relationiert sind. Zweitens können Assertionen Schlußfolgerungen veranlassen, die Veränderungen von beispielsweise Zuständen in der Wissensbasis zur Folge haben. In SYNDIKATE spielt beim augenblicklichen Stand der Implementation nur die erste Variante der konzeptuellen Interpretation eine Rolle. Dennoch sollen beide Möglichkeiten hier zur Darstellung kommen.

Betrachtet man die im Weltwissen modellierten Konzepte für lexikalische Einträge, die von der Wortklasse VERBAL subsumiert werden, kritisch, dann ist festzustellen, daß die Typrestriktionen, welche für konzeptuelle Relationen an diesen Verben aus den sprachlichen Verwendungszusammenhängen ableitbar sind, zu allgemein bleiben, um für das semantische Filtern noch brauchbar zu sein. Als Beispiel mag hier das transitive Verb „haben“

dienen, dessen Konzeptidentifizier im Weltwissen das Konzept POSSESS referenziert. In unserer Upper-Model-Ontologie ist der Restriktionstyp für das, was besessen wird (POSSESS-PATIENT), sehr allgemein, denn es kann sich um Eigenschaften, physikalische Teile oder auch um legales Eigentum handeln. Es ist unmittelbar einsichtig, daß der Typconstraint an der Relation sehr allgemein ist, insbesondere weil sehr viele Konzepte ein legales Eigentum sein können (Produkte, Ideen, Patente etc.). Dasselbe gilt auch für den Füller dessen, was besessen wird POSSESS-CO-PATIENT. Jedoch bestehen ganz strenge Restriktionen, wenn man die Füller der beiden Rollen zusammenbringt und deren Relationierbarkeit überprüft. Da sowohl ein Herzschrittmacher als Teil eines Patienten als auch eine Grafikkarte Teil eines Computers sein kann, ist die Zuordnung des Konzepts zum UPPER-MODEL der Wissensbasis legitimiert. Darüber hinaus sind auch die sehr generellen Typenrestriktionen nachvollziehbar. Dennoch läßt sich zweifelsfrei feststellen, daß – zumindest bis zum heutigen Tage – niemand eine Grafikkarte als Teil eines Patienten oder einen Herzschrittmacher Teil eines Computers vermuten würde. Mindestens ebenso einleuchtend ist, daß diese Form des Wissens vollständig von der syntaktischen Ebene abstrahiert und nur durch das Weltwissen repräsentiert werden kann.

Genau an dieser Stelle kommt die konzeptuelle Interpretation ins Spiel, indem sie die Constraints berechnet, die zwischen den Füllern der Instanzrollen bestehen. Dazu sind besondere Regeln zur konzeptuellen Interpretation definiert, die diese Prüfungen vornehmen, wenn ihr Bedingungsteil erfüllt ist. Im zitierten Beispiel würde also getestet, inwieweit eine Grafikkarte mit einem Computer relationierbar ist. Auch hier geht es wieder darum, zwei mit diesen Begriffen assoziierte Konzepte, i.e. GRAPHIC-CARD und COMPUTER-SYSTEM, über einen aus konzeptuellen Relationen bestehenden Pfad zu verbinden, um dadurch zusätzlich eine reichere inhaltliche Beschreibung in der Textwissensbasis zu erreichen. Im Negativbeispiel der Kombination von Herzschrittmacher und Computer würde ein derartiger Pfad nicht gefunden. Jedoch ist auch diese Suche Restriktionen unterworfen, die mit der Wortbedeutung von *“haben”* zusammenhängen. Betrachtet man die dargestellte Konstellation, dann wird deutlich, daß auch für die konzeptuelle Interpretation eine Spezialisierung der allgemeinen semantischen Interpretationsfunktion Verwendung finden kann. In der Funktion 4.6 findet sich die entsprechende Spezialisierung. Die Interpretationsfunktion abstrahiert vollständig von der syntaktischen Oberfläche, weil die Parameter zur Instantiierung ausschließlich von den Regeln zur konzeptuellen Interpretation bereitgestellt werden. Entsprechend übergibt eine konzeptuelle Interpretation zweie Konzepte C_1

und C_2 und eine Restriktionsmenge $R_{rule} \in 2^{\mathcal{R}}$, die am intervenierenden lexikalischen Knoten definiert ist, als Parameter an die Interpretationsfunktion, um gültige Relationspfade zu berechnen. Die Interpretationsfunktion ist Teil der Regeln für die konzeptuelle Interpretation.

$$si_{conc} : \begin{cases} \mathcal{F} \times 2^{\mathcal{R}} \times 2^{\mathcal{R}} \times \mathcal{F} \rightarrow 2^{R_{con}} \\ (C_1, R_{rule}, \emptyset, C_2) \mapsto \widetilde{R_{con}} \end{cases} \quad (4.6)$$

Die konzeptuelle Interpretation wird ausgeführt, sobald die zu interpretierende Struktur vollständig ist (der Bedingungsteil einer Regel erfüllt). In Abb. 4.9 findet sich die Interpretationsregel für POSSESS, das vom Konzept CAT-HAVING subsumiert wird, mit Bedingungs- und Aktionsteil. Wenn eine Instanz existiert, die den Konzepttyp CAT-HAVING besitzt, und zugleich die Instanzrelationen PATIENT und CO-PATIENT gefüllt sind, dann kommt es zum Triggern der konzeptuellen Interpretation. Dabei werden die Konzepttypen der Rollenfüller und die mit der Interpretationsregel assoziierte Restriktionsliste HAS-PART HAS-ATTRIBUTE HAS-LEGAL-PROPERTY als Parameter der Interpretationsfunktion übergeben. Für jeden der berechneten Relationspfade in REL wird eine entsprechende Assertion in die Textwissensbasis eingefügt.

Ergibt sich jedoch keine bezüglich des Weltwissens gültige konzeptuelle Interpretation, dann hat dies letztendlich Rückwirkung bis auf die Ebene des Parsing. Dies begründet sich dadurch, daß eine Kombination aus lokalen, semantisch interpretierten Teilgraphen für den Satz unter den Restriktionen der konzeptuellen Interpretation nicht kombinierbar ist. Damit muß die letzte, die Regel triggernde syntaktische Anbindung rückgängig gemacht werden und der Parser entsprechende Alternativen berechnen. D.h., die Interpretation macht nicht auf der Ebene eines einfachen Typchecks halt, sondern expandiert zunächst im Zuge der semantischen Interpretation einfache Attribut-Wert-Kombinationen auf wohlgeformte Pfade im Konzeptgraphen und prüft zusätzlich Abhängigkeiten zwischen Argumentstellen, wodurch reine Funktor-Argument-Strukturen eine deutliche Erweiterung finden.

Eine essentielle Frage, die sich an dieser Stelle stellt, ist die, wie das Regelwissen zur konzeptuellen Interpretation kodiert ist. Schließlich wäre es fatal, wenn die erreichte Beschreibungseconomie, welche alle bisher dargestellten Wissens Ebenen charakterisiert, dadurch ausgehebelt würde, daß zu jedem Verbkonzept – wie etwa POSSESS – je eine Regel zur konzeptuellen Interpretation existieren würde. Um dies zu vermeiden, benutzt auch die

```

EXISTS  $v, p, c$ :
   $v$  : CAT-HAVING  $\sqcap$ 
   $v$  PATIENT  $p$   $\sqcap$   $v$  CO-PATIENT  $c \implies$ 
  IF  $si_{conc}(type(p), \{HAS-PART HAS-ATTRIBUTE HAS-LEGAL-PROPERTY\}, \{\}, type(c)) \neq \emptyset$  THEN
    REL :=  $si_{conc}(type(p), \{HAS-PART HAS-ATTRIBUTE HAS-LEGAL-PROPERTY\}, \{\}, type(c))$ 
  TELL  $p$  REL  $c$  FORALL REL  $\in$  REL

```

Abbildung 4.9: Beispiel für eine Regel zur konzeptuellen Interpretation

konzeptuelle Interpretation bei der Abstraktion die Vorzüge des Vererbungsmechanismus, wie ihn die terminologische Logik bereit stellt. Diese Vererbung kommt auf zwei Ebenen ins Spiel. Zunächst existieren im domänenunabhängigen Teil der Wissensbasis drei Konzepte, ACTION, STATE oder PROCESS, für Verbkonzepte, die entweder eine Handlung, einen Zustand oder einen Prozeß darstellen. Das Konzept ACTION subsumiert die beiden in Satz (6) vorkommenden instantiierten Konzepte MANUFACTURE und SELL, STATE hingegen das Konzept POSSESS. Schließlich bündelt die zweite Ebene in den Konzepten VERB-CATEGORY und VERB-MODEL weitergehende Beschreibungen, die auf der Grundlage einer umfassenden Analyse der Verben des Deutschen (Ballmer & Brennenstuhl, 1986) entwickelt werden. Ballmer & Brennenstuhl (1986) bezogen in ihre empirische Studie 26.000 Verben ein und verwendeten eine Paraphrasetechnik, um den Aussagekern von Verben auf ein beschränktes Inventar von prädefinierten Paraphrasen zurückzuführen. Das Clustering dieser Paraphrasen zu Bedeutungsklassen und deren Einordnung in größere, skriptähnliche Handlungseinheiten führte zur Bildung von Kategorien bei der Gruppierung von Bedeutungen und zur Bildung von Modellen bei der Zusammenfassung von Verben zu Sequenzen. Die verwendeten Konzepte VERB-CATEGORY und VERB-MODEL setzen unmittelbar auf diesen Abstraktionen auf.

Während in den Konzepten ACTION, STATE und PROCESS über thematische Rollen die Argumentstellen für Rollenfüller an von Verben abgeleiteten Konzepten definiert sind, komprimiert das Konzept VERB-CATEGORY Ähnlichkeiten in der ontologischen Struktur. Das Subkonzept CAT-HAVING repräsentiert beispielsweise alle Aktionen und Zustände, in denen Objekte mit physikalischen Teilen, Eigenschaften oder rechtlichem Eigentum zusammengebracht werden. Dadurch daß die Interpretationsregeln an den Kategorien spezifiziert sind und nicht auf der Ebene einzelner Verbkonzepte, erreicht man auch auf dieser Ebene ein ontologisch sinnvolles Maß an Abstraktion, indem alle Verbkonzepte, die von CAT-

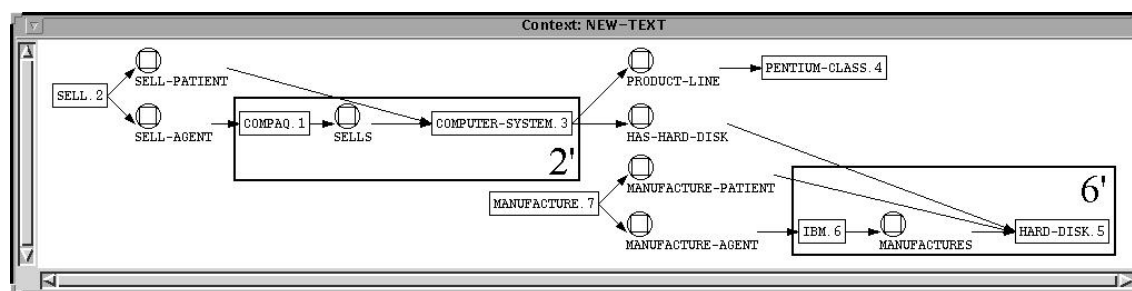


Abbildung 4.10: Konzeptuelle Interpretationen in Satz (6)

HAVING subsumiert werden, dieselbe Interpretationsregel nutzen. Zum aktuellen Zeitpunkt existieren ca. 30 konzeptuelle Interpretationsregeln, mit denen in etwa 200 Verbkonzepte abgedeckt sind.

Betrachten wir wieder unseren Satz (6), im Laufe der Analyse ergibt die *semantische* Interpretation des Satzanfangs “Compaq verkauft Computer” die zwei Assertionen in der Textwissensbasis: SELL.2 SELL-AGENT COMPAQ.1 und SELL.2 SELL-PATIENT COMPUTER-SYSTEM.3. Durch diese beiden Assertionen liegt in der Textwissensbasis eine aus zwei minimalen semantisch interpretierbaren Subgraphen bestehende Konstellation vor, durch die im vorliegenden Beispiel eine *konzeptuelle* Interpretation ausgelöst wird, d.h., der Bedingungsteil einer Regel zur konzeptuellen Interpretation ist erfüllt und der Aktionsteil wird ausgeführt. Dabei wird über die bekannte Pfadsuche berechnet, ob und wie COMPAQ.1 mit COMPUTER-SYSTEM.3 relationierbar ist. Ließe sich zwischen den Konzepttypen der beiden Instanzen kein gültiger Relationspfad berechnen, so müßte die angefragte syntaktische Relationierung, welche den Bedingungsteil der Regel für die konzeptuelle Interpretation erfüllbar macht, nämlich die in Abb. 4.1 in der Box 2 “*verkauft*” – *direct-object* – “*Computer*” zurückgenommen werden. Im positiven Fall, also der Relationierbarkeit der beiden Diskursobjekte COMPAQ.1 mit COMPUTER-SYSTEM.3, erfolgt eine Erweiterung der Textwissensbasis um die entsprechende Assertion. Gegeben die momentan in SYNDIKATE zur Verfügung stehende Wissensbasis terminiert der Pfadsuchalgorithmus mit der konzeptuellen Rolle SELLS als Ergebnis. Als Folgeschritt kommt die Assertion COMPAQ.1 SELLS COMPUTER-SYSTEM.3 in die Textwissensbasis (siehe Abb. 4.10). Um die bestehende Abhängigkeit der Berechnung der konzeptuellen Relation von der syntaktischen Relation in Abb. 4.1 Box 2 zu unterstreichen, ist in Abb. 4.10 die zu dieser konzeptuellen Interpretation gehörende Box mit 2' bezeichnet. Analog zum beschriebenen Fall führt die kon-

zeptuelle Interpretation der auf den Assertionen MANUFACTURE.7 MANUFACTURE-AGENT IBM.6 und MANUFACTURE.7 MANUFACTURE-PATIENT HARD-DISK.5 fußenden Konstellation zur Berechnung der Relation MANUFACTURES zwischen IBM.6 und HARD-DISK.5. Die dazugehörige Erweiterung der Textwissensbasis, die syntaktisch von der Anbindung des Relativsatzes an den Hauptsatz abhängt, umschreibt die Box 6'.

Die zweite angedeutete Art konzeptueller Interpretation, die Zustandsveränderungen von Diskursobjekten bewirkt oder auf tiefergehendem Reasoning basiert, ist in SYNDIKATE nicht erfaßt. In diesen Bereich gehören beispielsweise Textaussagen über den physikalischen Ort oder den Zustand eines Objekts. Aussagen dieser Art, wie sie in Texten enthalten sind, führen dazu, daß Assertionen ihre Gültigkeit verlieren, während andere neu in die Wissensbasis eingeführt werden müssen. Diese Schlußfolgerungen benötigen jedoch einerseits auf der Ebene des Knowledge Engineering ein sehr tiefes Modell über Handlungswissen sowie räumliches und temporales Schließen, andererseits auf der Ebene der Wissensrepräsentation ein System, welches in der Lage ist, diese Wissensebenen darzustellen und deren Implikationen zu berechnen. Beides ist in dieser Kombination mit ausreichender ontologischer Abdeckung nicht verfügbar.

Auf unsere Anwendungsdomänen bezogen dienen zur Illustration die nachfolgenden Beispiele. In der Informationstechnologie findet augenblicklich ein Konzentrationsprozeß statt, der in Fusionen bzw. Aufkäufen mündet. Aus dem Faktum eines Firmenkaufs – wie er mit den Komponenten zur semantischen und konzeptuellen Interpretation in SYNDIKATE ohne weiteres repräsentiert werden kann – ergeben sich eine ganze Reihe impliziter Konsequenzen, welche im Verkauf-Event enthalten sind und die nur auf der Grundlage eines detaillierten Event-Modells erschlossen werden können. Hierzu gehören die Transaktion der Kaufsumme vom Käufer auf die Eigner (ggf. in Form von Aktion), die Übertragung der Patentrechte und Markennamen vom Verkaufenden auf den Käufer, die Rechtsnachfolge seitens des Käufers für getätigte Geschäfte des Verkaufenden etc. Für die Medizin-Domäne ließe sich als Beispiel das Entfernen eines Blinddarms anführen. Im Anschluß an diese Handlung gilt nicht mehr, daß der betroffene Patient einen Blinddarm hat, wohingegen vor dem Eingriff diese Aussage noch Gültigkeit besitzt.

Mit diesen Beispielen wird offensichtlich, daß zum einen zur Repräsentation dieser temporalen Bezüge (Gültigkeit von Aussagen bezüglich eines Zeitpunkts t) unsere Form der Wissensdarstellung unzureichend ist und zum anderen Art und Umfang des bereitgestellten Wissens den hier anzutreffenden Reasoninganforderungen nicht genügen. Damit sind die

Grenzen für die Wissensextraktionsfähigkeit von SYNDIKATE gesteckt. Im Kern wird das in den Texten durch syntaktische Strukturen und Diskursreferenten explizit gemachte Wissen in eine an den Abhängigkeitsstrukturen angelehnte inhaltliche Repräsentation überführt. Der Überprüfung von Wohlgeformtheitskriterien bei der semantischen Interpretation wird noch eine Form der konzeptuellen Interpretation zur Seite gestellt, die Integritätsbedingungen zwischen Füllern von Instanzrollen an derselben Instanz kontrolliert. Obwohl der Wissensextraktionsprozeß nicht weitergeht, sei angemerkt, daß weitergehende Berechnungen zum tieferen Verständnis eines Texts auf der gewonnenen Inhaltsrepräsentation aufsetzen können.

4.2 Ansätze zur semantischen Interpretation

In Kapitel 2.2 kamen bereits schon Ansätze zur semantischen Interpretation zur Darstellung, wobei das Zusammenspiel von Syntax und Semantik sowie die Anzahl der implementierten Regeln mit angesprochen wurden. In diesem Kapitel soll dieses Thema nochmals unter einer formal-linguistischen Perspektive kurz aufgegriffen werden. Dabei werden zwei Aspekte berührt. Zuerst geht es um die Anzahl der Regeln, die für die semantische Interpretation benötigt werden, und anschließend darum, wie die semantische Interpretation syntaktische Strukturen auf inhaltliche Repräsentationen abbildet.

Eine Möglichkeit zur semantischen Interpretation besteht darin, mit jeder syntaktischen Regel eine oder gar mehrere semantische Interpretationsregel(n) zu verbinden (van Eijck & Moore, 1992). Mit Ausnahme der Beschränkungen, die sich aus der gewählten semantischen Theorie ergeben, lassen sich dabei semantische Regeln weitgehend frei und ohne Einschränkung formulieren. In diesem Ansatz fehlen Abstraktionsmechanismen, die der Beschreibungsökonomie förderlich sind, abgesehen von einer taxonomischen Ordnung der semantischen Bezeichner (Creary & Pollard, 1985), nahezu vollständig. Dementsprechend benötigen Ansätze, die nur wenig abstrahieren, in realistischen Anwendungsszenarien bei nur kleinen Domänen sehr schnell eine große Regelmenge. Das bereits in Kapitel 2.2 angeführte System von Bean et al. (1998) erschließt die Interpretation ausschließlich räumlicher Beziehungen aus Texten über den Thorax mit bereits mehreren hundert Regeln. Eine Begrenzung der Regelanzahl ist aber notwendig, um Skalierbarkeit, Portierbarkeit und Wartbarkeit zu gewährleisten. Auf dieser Ebene bringt das schlanke Modell zur semantischen Interpretation in SYNDIKATE eine deutliche Verbesserung mit sich.

Als weiteres konzeptionelles Merkmal von SYNDIKATE gibt es die Modularisierung der Wissenssebenen, die gerade bei dem schlanken Modell zur semantischen Interpretation eine wichtige Rolle spielt. Dabei laufen die semantischen Berechnungen zu syntaktischen Strukturen direkt in der Textwissensbasis ab. Ähnlich verfahren Sondheimer et al. (1984) und Hirst (1988), die syntaktische Strukturen ebenfalls ohne intermediäre Repräsentationsschicht auf konzeptuelle Strukturen abbilden. Eine weitere Analogie zu SYNDIKATE besteht darin, daß das Domänenwissen ebenfalls in einer KL-ONE-artigen terminologischen Sprache vorliegt. In der Wissensrepräsentation finden sich demnach Abstraktionen in Form von taxonomischen Beziehungen und Vererbung. Jedoch sind die Übersetzungsregeln von der syntaktischen auf die konzeptuelle Ebene bei Sondheimer et al. (1984) an die einzelnen Rollen von Konzeptdefinitionen gekoppelt und spiegeln damit die Idiosynkrasien jeder einzelnen Rollenbeschreibung wider. Nur im Fall von Konzeptspezialisierungen kommt die beschreibungsökonomisch wirksame Vererbung von Regeln zum tragen. In einer ähnlichen Art und Weise benutzt Hirst (1988) Typisierungen nur auf der Ebene von Konzeptbeschreibungen. In SYNDIKATE finden sich die notwendigen Abstraktionen auf *allen* Wissenssebenen: auf der Lexikon/Parser-Seite durch Wortklassenhierarchien, in den Wissensbasen durch Konzepthierarchien mit Eigenschaftsvererbung und schließlich auf der semantischen Ebene durch die Spezialisierung eines generischen Interpretationsschemas.

Bei Charniak & Goldman (1988) und Jacobs (1991) finden sich bereits spezifische semantische Regeln innerhalb einer Vererbungshierarchie. Dadurch kommen sie der in SYNDIKATE verwendeten schlanken Methode insgesamt am nächsten. Allerdings koppelt Jacobs (1991) die Spezifikation syntaktischer Relationen vollständig an die konzeptuellen Rollen, wodurch sich – wie bei den in Kapitel 2.2 angesprochenen semantischen Grammatiken – eine Vermischung der Wissenssebenen einstellt, die eine Modularisierung der Wissensquellen aufhebt. Charniak & Goldman (1988) stellen eine generelle Regel für Kasusrollen auf, die dem Interpretationsschema für *direct-linkage* ähnelt. Jedoch bleibt das Interpretationsschema für *direct-linkage* allgemeiner, da sich mit diesem neben allen Kasusrollen auch Genitivattribute und Adjektive interpretieren lassen. Deshalb müssen Charniak & Goldman (1988) für jedes syntaktische Phänomen eine eigene Regel spezifizieren. Aufsattpunkt für semantische Regeln sind syntaktische Regeln und nicht wie in SYNDIKATE abstraktere Konfigurationen in einem Abhängigkeitsgraphen.

Insgesamt begünstigt die strukturelle Ähnlichkeit der Abhängigkeitsgraphen mit den durch die semantische Interpretation berechneten konzeptuellen Graphen die Abstraktion (Ha-

jicova, 1987). In der *direct-linkage*-Konstellation unterstreicht die statische Interpretation nochmals diesen engen Bezug, der sich darin ausdrückt, daß sich binäre Dependenzrelationen auf eine Menge von binären konzeptuellen Interpretationen abbilden lassen. Trotz dieser strukturellen Ähnlichkeit gibt es wenige Arbeiten darüber, wie Dependenzstrukturen adäquat semantisch zu interpretieren sind (vgl. Milward (1992) und Lombardo et al. (1998)). Im Gegensatz dazu ist das Thema in konstituentenbasierten Ansätzen vergleichsweise gut erschlossen (z.B.: Moore (1989), Dalrymple et al. (1993) oder Wedekind & Kaplan (1993)).

Die Frage, inwieweit der vorgestellte Ansatz auch auf andere Grammatikansätze übertragbar ist, kann im Rahmen dieser Arbeit nicht geklärt werden. Prinzipiell wäre es indessen denkbar, beispielsweise in einem konstituentenbasierten Ansatz die Regeln von den Konstituenten an die inneren Knoten durchzupropagieren, wie dies in den Arbeiten von Pereira & Pollack (1991) und Milward (1995) geschieht.

Im Gegensatz zu den logik-basierten Ansätzen zum Textverstehen kann die semantische Komponente von SYNDIKATE Quantifikation nicht interpretieren (vgl. zu diesem Thema z.B. Charniak & Goldman (1988), Moore (1989) und Pereira & Pollack (1991)). Hierin liegt aus formal-linguistischer Sicht sicherlich eines der größten Defizite. Auch die mit der Quantifikation verbundene Repräsentation von Skopusambiguitäten entfällt in SYNDIKATE (zum Thema Skopusambiguitäten, vgl. z.B. Carpenter (1997)). Die Gründe hierfür liegen unter anderem im verwendeten Wissensrepräsentationsformalismus, der für diese beiden wichtigen Sprachphänomene keine adäquaten Ausdrucksmöglichkeiten bietet. Im Gegensatz dazu verwenden Lombardo et al. (1998), Milward (1994) und Pereira (1990) Formalismen zur semantischen Interpretation die auf dem (getypten) λ -Kalkül aufsetzen, wodurch sich durchgehend quantifizierte semantische Repräsentationen für Eingabesätze ergeben.

Der genaue Zusammenhang in SYNDIKATE zwischen der Ausdrucksmächtigkeit des eingesetzten formalen System einerseits und den konkreten Anforderungen aus Textverstehen und Wissensrepräsentation andererseits kommt ausführlich in Kapitel 5 zur Darstellung, in dem auch die Begründung für die beschriebenen Defizite angeführt wird.

Als Fazit dieser kurzen Übersicht über Ansätze zur semantischen Interpretation kann eine Reihe von Punkten festgehalten werden. Grundsätzlich erzielen sprachtheoretisch orientierte Arbeiten reichhaltigere Interpretationsergebnisse als SYNDIKATE. Jedoch findet man i.a. keine Aussagen über die Leistungsfähigkeit in realistischen Anwendungen (was

auch nicht in deren Intention liegt). Außerdem bleibt der Übergang zwischen den einzelnen semantischen Ableitungsschritten verborgen. Es wird nicht gezeigt, wie die Inferenzschritte das notwendige Domänenwissen bei der Ableitung abrufen und wie das Domänenwissen in den Interpretationsprozeß integriert wird. Noch weniger erfährt man über Umfang und Abdeckung der Wissensquellen, die die semantische Interpretation unterstützen. Jedoch kann sich die Menge an Interpretationsregeln schnell in den Bereich mehrerer hundert bis tausend hochschwingen, wenn auf spezifischen syntaktischen Strukturen gearbeitet wird.¹⁰ Im Gegensatz dazu bietet SYNDIKATE ein transparentes und schlankes Verfahren für die semantische Interpretation an, das an authentischen Freitexten evaluiert wurde (siehe Kap. 6). Zieht man indessen statt formal-linguistischer Ansätze Informationsextraktionssysteme zum Vergleich heran, dann ist SYNDIKATE in der Lage, mehr Informationen zu extrahieren, da es keine Beschränkung auf vordefinierte Schemata gibt.

4.3 Das Ambiguitätsproblem bei der Wissensextraktion aus Texten

Jedes System, das natürliche Sprache verarbeitet, muß eine formale wie auch implementatorische Lösung für die Behandlung von Ambiguitäten vorhalten. Mehrdeutigkeiten treten dabei auf verschiedenen Beschreibungsebenen auf. Die Erzeugung einer validen Textwissensbasis für einen Eingabetext hängt maßgeblich davon ab, wie Ambiguitäten in einem Wissensextraktionssystem verarbeitet und reduziert werden.

In diesem Unterkapitel soll daher erläutert werden, wie im Textverstehenssystem SYNDIKATE mit den unterschiedlichen Erscheinungsformen von Ambiguitäten umgegangen wird. Zu diesem Zweck erfolgt zunächst in Kap. 4.3.1 eine Aufstellung der Ambiguitätsphänomene, die während der Analyse eines Texts auftreten. Anschließend wird der Kontextbegriff unter zwei unterschiedlichen Perspektiven aufgegriffen, um die Erweiterungen der in Kap. 3.2.2 eingeführten terminologischen Logik um einen Kontextmechanismus zu motivieren. Das nachfolgende Kapitel 4.3.3 behandelt die erweiterten prozeduralen Grundlagen der Interaktion von Parser, Semantik und Weltwissen bei der Erzeugung ei-

¹⁰Interessant in diesem Zusammenhang sind auch die *meaning postulates* im Sinne einer Montague-Grammar (Montague, 1974), die eine sehr detaillierte Regelkodierung erforderlich machen (siehe in etwa die Arbeit von Paredes-Frigoletti & Strube (1996)).

ner Textwissensbasis. Abgerundet wird die Ambiguitätsthematik mit Beispielen für die jeweiligen Phänomenklassen, um die Flexibilität und Vorteile des eingeführten Kontextmechanismus' zur Darstellung zu bringen.

Der Schwerpunkt der Ausführungen liegt in der Präsentation der zugrundeliegenden Repräsentationsmechanismen und deren Integration in den Textanalyseprozeß. Eine vollständige Formalisierung konnte im Zuge dieses Promotionsvorhabens nicht geleistet werden.

4.3.1 Ambiguitäten und das Textverstehenssystem SYNDIKATE

Wie in Kap. 3 bereits aufgezeigt, bearbeitet der Parsing-Kern von SYNDIKATE die Eingabe inkrementell. Darüber hinaus verfügt der Parser über keine globale Steuerung der Berechnungsprozesse, vielmehr werden Verträglichkeitsbedingungen für eine Dependenzrelation lokal verrechnet. Aus diesem Grund entstehen während der Satzanalyse eine Reihe von lokalen Ambiguitäten (vgl. Kap. 4.3.4), deren Anzahl sich durch eine andere Analysestrategie reduzieren ließe. Andererseits erreicht man gerade durch die Lokalität der Berechnungen, also durch die relative Unabhängigkeit vom globalen syntaktischen Kontext, die notwendige Robustheit gegenüber un-, extra- und paragrammatischer Eingabe.

Aufgrund des modularen Aufbaus von SYNDIKATE lassen sich die verschiedenen Ambiguitätsphänomene gut isolieren und beschreiben.¹¹ Die nachfolgende Aufstellung enthält die Paradigmen von Ambiguitäten, wie sie SYNDIKATE zu verarbeiten und zu repräsentieren in der Lage ist:

- *Lexikalische* Ambiguitäten resultieren einerseits aus den unterschiedlichen Wortklassenzugehörigkeiten einer Wortoberfläche im Text (*part-of-speech*: z.B. "haben" als temporales Auxiliar oder als transitives Verb), andererseits aus der Mehrdeutigkeit einer Wortoberfläche im Text bezüglich seiner Bedeutung (*Polysemie*: z.B. "Bus" als Fahrzeug oder als Hardwarekomponente). Die beiden angeführten Formen lexikalischer Ambiguität können kombiniert auftreten (vgl. 4.3.4).
- *Strukturelle* Ambiguitäten ergeben sich dadurch, daß sich eine Phrase oder ein Wortaktor an Valenzen verschiedener im Satz enthaltener lexikalische Köpfe anbin-

¹¹Unabhängig von der thematischen Fokussierung auf das Textverstehenssystem SYNDIKATE sind die aufgeführten Fälle allgemein gültig.

den kann. Das prominenteste Beispiel für strukturelle Ambiguitäten stellt die PP-Anbindung dar. Eine Präpositionalphrase kann sowohl als Verbergänzung als auch als Modifikator zu einem Nomen kategorisiert werden. Aufgrund dieser syntaktisch gleichwertigen Anbindungsmöglichkeiten entstehen strukturell divergierende Dependenzgraphen.¹²

- *Semantische* Ambiguitäten ergeben sich aus dem Interpretationsprozeß in der Textwissensbasis, wenn die Suche nach validen Pfaden in der Domänenwissenbasis mehr als nur eine konzeptuelle Relationierung berechnet. Die Interpretation des Genitivattributs in der Phrase “*Das Speichern des Programms*” erlaubt zum Beispiel – gegeben die IT-Wissenbasis in ITSYNDIKATE –, daß ein “*Programm*” sowohl AGENT als auch PATIENT eines Speicherungsereignisses sein kann. Da Genitivattribute an Verbnominalisierungen eine restringierte Suche im Konzeptgraphen ausführen (vgl. 4.1.1), die sich auf die zu *subject* und *direct-object* kompatiblen konzeptuellen Relationen beschränkt, führt die Interpretation zu zwei validen Relationierungen (i.e. STORE-AGENT und STORE-PATIENT), die adäquat repräsentiert werden müssen.
- *Textuelle* Ambiguitäten folgen aus der Resolution satzübergreifender referentieller Bezüge zwischen Diskursobjekten, die durch anaphorische Ausdrücke angezeigt werden.¹³

Die Behandlung von Ambiguitäten, welche auf der Wort-, der Phrasen- und Satzebene sowie der Textebene anzutreffen sind, setzt auf zwei Ebenen an. Erstens muß ein formaler Rahmen geschaffen werden, in dem die inhaltliche Repräsentation der unterschiedlichen Lesarten konsistent abzubilden ist. Zweitens hat im Laufe des Analyseprozesses kontinuierlich eine Disambiguierung stattzufinden, die die Anzahl der Lesarten möglichst klein hält und ausschließlich valide Lesarten weiterverarbeitet. Die Disambiguierung setzt hierbei auf der Interaktion und Integration lexikalischen, grammatikalischen, textuellen, semantischen wie auch konzeptuellen Wissens auf.

Die innerhalb von SYNDIKATE realisierte Lösung für die beiden Problemschichten verwendet einen Kontextmechanismus, durch den einerseits Ambiguitäten repräsentierbar

¹²Im Falle der Verwendung eines anderen Grammatikformalismus¹ werden die strukturellen Mehrdeutigkeiten entsprechend durch andere Grammatikstrukturen repräsentiert.

¹³Beim gegenwärtigen Stand der Implementation in SYNDIKATE treten keine Ambiguitäten auf dieser Beschreibungsebene auf, weil erfolgreich resolvierte Antezedenten aus der Centering-Liste gelöscht werden.

sind und andererseits die Disambiguierung von Lesarten unmittelbar unterstützt wird. Der Begriff “*Kontext*” kennt in der wissenschaftlichen Diskussion zwei Verwendungszusammenhänge (vgl. hierzu (Hirst, 1997)).¹⁴ Zum einen betrachtet man Kontexte als formale Objekte und untersucht deren Eigenschaften unter der Perspektive der Wissensrepräsentation (McCarthy, 1993). Zum anderen begreift man Kontext(e) in einer linguistisch-sprachorientierten Perspektive als die rein sprachlichen und pragmatischen Bedingungen, unter denen eine Äußerung zu interpretieren ist (Ferrari, 1997). Insbesondere der pragmatische Rahmen, d.h. der nicht-sprachliche oder metasprachliche Kontext einer Äußerung dient in Alltagssituationen der Disambiguierung und verhindert damit Fehlinterpretationen.

Im Szenario des maschinellen Textverstehens steht ein pragmatischer Rahmen indessen nicht zur Verfügung. Der sprachliche Kontext, in dem sich ein Text entfaltet, ist alleine durch den Textinhalt selbst gegeben. Zur Reduktion von Ambiguitäten können daher nur Informationen aus dem bereits analysierten Vortext herangezogen werden. Darüber hinaus steht indessen noch das a priori bereitgestellte Domänenwissen zur Verfügung, mit dem geprüft wird, ob syntaktische Lesarten im Zuge der semantischen Interpretation auch vom konzeptuellen Wissen sanktioniert werden (“*semantic filtering*”, vgl. Bonnema et al. (1997)). Im Gegensatz zu der sich im Analyseprozeß dynamisch entfaltenden Textwissensbasis bleibt das konzeptuelle Hintergrundwissen statisch. In Anlehnung an Zarri (1995) könnte man hier auch von dynamischem und statischem Kontext sprechen.

4.3.2 Erweiterung der terminologischen Logik um einen uniformen Kontextmechanismus

Ich stelle nun vor, wie im Textverstehenssystem SYNDIKATE ein an die terminologische Logik adaptierter, jedoch eingeschränkter Kontextmechanismus Verwendung findet. Kontexte sind formale Objekte (vgl. hierzu McCarthy & Buvač (1994)), die in unserem Ansatz eine Menge an terminologischen und assertionalen Axiomen einschließen. Mit Hilfe dieser Kontexte sind unterschiedliche Hypothesen über die während der Textinterpretation extrahierbaren Aussagen zu repräsentieren. Alternative, jedoch logisch konsistente und

¹⁴Die Doppeldeutigkeit des Kontext-Begriffs läßt sich leider für die nachfolgenden Ausführungen nicht ausräumen. Zur Verdeutlichung des Verwendungszusammenhangs sind Kontexte unter Wissensrepräsentationsaspekten als formale Kontexte markiert.

syntaktisch sanktionierte Bedeutungsrepräsentationen eines Eingabetextes in Form von assertionalen Axiomen sind als disjunkte Hypothesen über den Textinhalt in separaten Kontexten gekapselt, d.h., die Veroderung der in den Kontexten enthaltenen assertionalen Axiome stellt das Interpretationsergebnis dar (Buvač, 1995). Damit lassen sich die beiden Bedeutungen des Begriffs “Kontext” auseinanderhalten: *Formale Kontexte* dienen als Container, um valide Aussagen über den Bedeutungsgehalt von Sätzen in dem durch den Text selbst gegebenen *sprachlichen Kontext* aufzunehmen.

Zunächst sei die Menge aller Kontexte $\mathcal{H}$ gegeben.¹⁵ Zur Einführung einer Vererbungshierarchie über Kontexte sei die transitive und reflexive Relation *subcontext-of* $\subset \mathcal{H} \times \mathcal{H}$ definiert. Um die in einem Kontext $h \in \mathcal{H}$ enthaltene T-Box bzw. A-Box zu kennzeichnen, findet die Schreibweise $\mathcal{T}_h$ bzw. $\mathcal{A}_h$ Verwendung. Sind zwei Kontexte $h_1 \in \mathcal{H}$ und $h_2 \in \mathcal{H}$ über *subcontext-of* miteinander relationiert, dann gilt zwischen den jeweiligen T-Boxen und A-Boxen eine der Vererbung zugrundeliegende Teilmengenbeziehung. Das heißt, daß alle terminologischen und assertionalen Axiome, $\mathcal{T}_{h_1}$ und $\mathcal{A}_{h_1}$ eines Elternkontexts h_1 auch in dessen Kindkontext h_2 enthalten, also Teilmenge von $\mathcal{T}_{h_2}$ und $\mathcal{A}_{h_2}$, sind. Diese informell gegebene Definition erlaubt prinzipiell multiple Vererbung, wodurch Kontexthierarchien in Form von Graphen (von nun an als *Kontextgraphen* bezeichnet) möglich sind. Jedoch gelten für diese Kontexthierarchien eine Reihe von Bedingungen:

- Die Menge der terminologischen und assertionalen Axiomen wächst in der Vererbungshierarchie monoton, was unmittelbar aus den Teilmengenbeziehungen hervorgeht. Damit verbietet sich insbesondere die Zurücknahme von Assertionen.
- Auf der Ebene der terminologischen Axiome gilt darüber hinaus, daß eine kontextabhängige Konzeptdefinition untersagt ist. D.h., wenn ein Konzeptidentifizier in einem Kontext durch ein terminologisches Axiom definiert ist, dann gilt diese Definition in allen anderen Kontexten, in denen dieser Konzeptidentifizier – sei es aufgrund von Vererbung oder per Definition – ebenfalls auftritt.

Durch diese Einschränkung kann eine T-Box in einem beliebigen Kontext innerhalb

¹⁵Das Gewicht liegt bei dieser Darstellung nicht auf dem formalen Beitrag, da hier bereits grundlegende Vorarbeiten existieren (Buvač & Mason, 1993; 1994; Buvač et al., 1995; Buvač, 1995), sondern auf der konkreten Anwendung im Rahmen eines Textverstehenssystems. Eine Übersicht zu Arbeiten über Kontexte findet sich in Akman & Surav (1996). Ansätze zur formalen Definition von Kontexten behandelt Dichev (1997).

des Kontextgraphen zwar durch Hinzufügung neuer Konzept- und Rollendefinitionen erweitert werden, eine kontextabhängige Definition desselben Konzeptsymbols hingegen entfällt. In SYNDiKATE findet sich diese Form der Erweiterung im vom System a priori bereitgestellten terminologischen Wissen $\mathcal{T}_{it}$ (für die Informationstechnologie) und $\mathcal{T}_{med}$ (für die Medizin), da diese als T-Boxen der Kontexte $\mathcal{KB}_{it}$, $\mathcal{KB}_{med} \in \mathcal{H}$ die Terminologie des UPPER-MODEL erweitern (vgl. hierzu auch Abb. 3.7). Schneidet man jedoch die Menge der Konzept- und Rollenbezeichner zweier Kontexte, dann gilt für alle in der Schnittmenge enthaltenen Symbole, daß sie jeweils durch denselben terminologischen Ausdruck definiert sein müssen. Damit ist auch ein “Überschreiben” von Definitionen ausgeschlossen.

Im Gegensatz dazu dient die kontextspezifische Interpretation von Individuensymbolen der Repräsentation eines Textinhalts. Folglich können Individuen oder Tupel von Individuen in Abhängigkeit von Kontexten in der Extension unterschiedlicher Konzepte bzw. Relationen liegen.

- Per Konvention existiert in allen Kontexten dieselbe Menge von Instanzsymbolen, wodurch die Domäne Δ (das Diskursuniversum) für alle Kontexte identisch ist.

Da im Gegensatz zu terminologischen Axiomen auf der assertionalen Ebene für einen Instanzbezeichner kontextabhängige Axiome abgesetzt werden können, wird die modelltheoretische Semantik assertionaler Axiome in Anlehnung an Schnattinger & Hahn (1998) um eine kontextabhängige Interpretation erweitert. Da die T-Box im gegebenen Anwendungsszenario unveränderlich ist, fällt sie aus den weiteren Betrachtungen heraus. In Tabelle 4.1 finden sich diese Erweiterungen wieder. Assertionen sind an Kontexte gebunden, was syntaktisch durch die Klammerung mit einem Kontextsymbol als Index markiert wird. Sei α ein terminologisches Axiom und $h \in \mathcal{H}$ ein Kontext, dann bezeichnet $(\alpha)_h$, daß $\alpha \in \mathcal{A}_h$. Die Interpretation $\mathcal{I}$ der Instanzbezeichner erfolgt nun in Abhängigkeit zu einem Kontext h aus der Menge der Kontexte $\mathcal{H}$, in dem sich ein assertionales Axiom befindet, und wird dementsprechend im folgenden als $\mathcal{I}_h$ bezeichnet. D.h., die Interpretationsfunktion $\llbracket \cdot \rrbracket^{\mathcal{I}}$ ergibt bei gleicher Domäne Δ unterschiedliche Belegungen in Abhängigkeit vom Kontext. Entsprechend findet sich in Tabelle 4.1 eine Beschreibung der erweiterten Semantik für die Zugehörigkeit einer Instanz a zu einem Konzept C bzw. die Relationierung zweier Instanzen a und b über eine Relation R . Wird ein Kontext h beispielsweise durch (SELL.2

Erweiterte assertionale Axiome	Semantik
$(a : C)_h$	$a^{\mathcal{I}_h} \in C^{\mathcal{I}_h}$
$(a R b)_h$	$(a^{\mathcal{I}_h}, b^{\mathcal{I}_h}) \in R^{\mathcal{I}_h}$

Tabelle 4.1: Syntax und Semantik von kontextabhängigen assertionalen Axiomen

: $\text{SELL})_h$ um das assertionale Axiom ($\text{SELL.2} : \text{SELL}$) erweitert, dann liegt SELL.2 in der Interpretation $\mathcal{I}_h$ des Konzepts SELL bezüglich des Kontexts h , also $\text{SELL.2}^{\mathcal{I}_h} \in \text{SELL}^{\mathcal{I}_h}$.

Die um einen Kontextmechanismus erweiterte terminologischen Logik findet im Textanalyseprozeß Verwendung, um inkonsistente semantische Interpretationen auszufiltern und wirkt damit zudem auf den Suchraum der syntaktischen Berechnungen zurück (*semantic filtering*, vgl. Kap. 3.2.4). Eine T-Box ist konsistent, wenn eine Interpretation $\mathcal{I}$ existiert, die alle Axiome der Terminologie erfüllt. D.h., es gibt mindestens eine Interpretation, unter der jedem Konzept eine mindestens einelementige Extension zugewiesen werden kann (Baader et al., 1992). Diese Interpretation ist dann ein Modell für die T-Box. Umgekehrt ist eine Konzeptbeschreibung C inkonsistent, wenn eine T-Box impliziert, daß die Extension eines Konzepts durch die leere Menge gegeben ist (formal: $\mathcal{T} \models C^{\mathcal{I}} = \emptyset$) (Donini et al., 1996); eine derartige T-Box hat dann kein Modell. Für die nachfolgende Betrachtung der assertionalen Ebene gelte ohne Einschränkung, daß $\mathcal{T}_{upper}$, $\mathcal{T}_{it}$ und $\mathcal{T}_{med}$ sowie deren Vereinigung konsistente Terminologien darstellen. Aufgrund der Vereinbarung, daß die T-Box während der Textanalyse innerhalb der Kontexthierarchie unveränderlich bleibt und kontextabhängige Konzeptdefinitionen untersagt sind, erübrigt sich eine kontextabhängige Betrachtung terminologischer Axiome.

Eine A-Box ist bezüglich einer T-Box genau dann konsistent, wenn sie ein Modell hat, d.h., die Interpretation weist jedem Instanzsymbol genau ein Element der Domäne Δ zu. Im Gegensatz dazu ist eine A-Box inkonsistent, wenn die Interpretation $\mathcal{I}$, gegeben eine T-Box mindestens einem Instanzsymbol a , kein Element der Domäne Δ zuweisen kann. Dieser Zusammenhang spielt deshalb beim Textverstehen eine Rolle, weil durch das Hinzufügen von Assertionen inkonsistente Interpretationskontexte, also solche mit unerfüllbaren assertionalen Axiomen, entstehen können. Diese sind dann von einer weiteren Verarbeitung bei der Textanalyse ausgeschlossen. D.h., sie werden nicht weiterverfolgt.

Für die Beschreibung der Interaktion des Parsers mit der Textwissensbasis werden weiterhin die Relation *direct-subcontext-of* $\subset \mathcal{H} \times \mathcal{H}$ und das Prädikat *terminalContext*(h)

eingeführt. Die Relation *direct-subcontext-of* ist dadurch definiert, daß ein Kontext h_1 mit einem Kontext h_3 über genau eine *subcontext-of* Relation miteinander verbunden ist. Es existiert folglich kein Kontext h_2 , der Kindkontext von h_1 und zugleich Elternkontext von h_3 ist. Analog zu den bereits eingeführten Relationen *subcontext-of* und *direct-subcontext-of* sind deren Inverse *supercontext-of* und *direct-supercontext-of* definiert, ohne daß diese formal explizit eingeführt werden.

Das Prädikat *terminalContext(h)* für einen Kontext $h \in \mathcal{H}$ ist erfüllt, gdw. kein Kontext h' existiert, der über *subcontext-of* mit h relationiert ist.

4.3.3 Integration des Kontextmechanismus in SYNDIKATE

Im Anschluß an die Klärung der formalen Voraussetzungen für den Kontextmechanismus befaßt sich dieses Unterkapitel mit den konkreten Zusammenhängen, wie das Textverstehenssystem SYNDIKATE die eingeführten Ambiguitätsphänomene mit Hilfe von Kontexten verarbeitet. Generell lassen sich vier Ebenen unterscheiden, in denen Kontexte Verwendung finden. Allerdings sind nur die letzten drei für die Ambiguitätsbehandlung relevant:

- **Wissensmodellierung.** Auf der Ebene der Wissensmodellierung kommt die Designphilosophie der Modularisierung und der Abstraktion dadurch zum Ausdruck, daß die domänenspezifischen Wissensbasen für Medizin und Informationstechnologie Kindkontexte des domänenunabhängigen “Upper-Models” sind.
- **Initialisierung.** Während der Initialisierungsphase werden auf der Syntaxebene die Wortaktoren und innerhalb der Textwissensbasis die Diskursentitäten instantiiert. Folglich sind syntaxseitig Wortart-Ambiguitäten zu repräsentieren und in der Textwissensbasis konzeptuelle Mehrdeutigkeitsphänomene zu behandeln .
- **Semantische Interpretation.** Mehrdeutigkeiten treten im Zusammenhang mit der semantischen Interpretation in drei Fällen auf. Zunächst sind die differierenden inhaltlichen Interpretationen *syntaktischer* Ambiguitäten zu repräsentieren. Weiterhin kann die Pfadsuche bei der *semantischen* Interpretation mehr als nur einen gültigen Relationspfad als interpretatorische Alternativen berechnen. Schließlich können Ambiguitäten auch bei der *satzübergreifenden Referenzauflösung* auftreten.
- **Selektion.** Trotz der Unvollständigkeit des Parsers gibt es i.a. für einen Eingabesatz am Satzende eine Reihe von alternativen Interpretationen. Aus Komplexitätsgründen

können indessen nicht alle Lesarten weiterverfolgt werden, vielmehr erfolgt am Satzende eine Selektion der besten Lesarten für die weitere Berechnung.

Wie bereits im Kapitel zur Systemarchitektur erläutert, besteht zu Beginn der Analyse eines Eingabetextes die initiale Textwissensbasis aus einer exakten Kopie der zur Textklasse gehörenden Domänenwissensbasis, je nachdem ob es sich um einen Text aus der Informationstechnologie oder um einen Text aus der Medizin handelt. Diese Kopie läßt sich nunmehr auf der Grundlage der definierten Vererbung von assertionalen und terminologischen Axiomen zwischen Kontexten leicht erzeugen, indem ein direkter Subkontext zu $\mathcal{KB}_{it}$ bzw. $\mathcal{KB}_{med}$ angelegt wird. Die zu einem Eingabetext gehörende initiale Textwissensbasis trägt im folgenden per Konvention grundsätzlich die Bezeichnung NEWTEXT, ohne dabei zu berücksichtigen, daß natürlich zu jedem Text eine eigene Textwissensbasis angelegt werden muß. Da bis zum Abschluß der Verarbeitung eines Eingabetextes ausschließlich Assertionen in die Textwissensbasis eingestellt werden, bleibt die Terminologie (die T-Box) in der Textwissensbasis grundsätzlich unverändert, nämlich $\mathcal{T}_{it}$ bzw. $\mathcal{T}_{med}$.¹⁶ Aufgrund der konstanten T-Boxen beschränken sich die nachfolgenden Ausführungen ausschließliche auf die assertionale Ebene, um die unterschiedlichen Ebenen der Ambiguität, die während des Textanalyseprozesses auftreten, über den Kontextmechanismus zu repräsentieren.

Um zu gewährleisten, daß das über die Instanzbezeichner eingeführte Diskursuniversum eines Texts für alle Kontexte identisch ist, wird jeder Instanzbezeichner (als Platzhalter für eine Diskursentität) in der durch den Kontext NEWTEXT repräsentierten initialen Wissensbasis ungetypt eingeführt. Da in SYNDIKATE alle Interpretationskontexte für Interpretationsalternativen direkte oder indirekte Subkontexte der initialen Textwissensbasis NEWTEXT sind, stellt der Vererbungsmechanismus sicher, daß die Individuenkonstanten gleichermaßen in allen Kontexten beinhaltet sind, d.h. Δ ist bezüglich der Kontexte invariabel.

Sei a der Bezeichner einer einzuführenden Instanz, dann wird $\mathcal{A}_{NewText}$ um die Assertion $\alpha := (a : \top)$ erweitert, d.h. $\mathcal{A}_{NewText} := \mathcal{A}_{NewText} \cup \{\alpha\}$. Obwohl für alle Kontexte die gleiche Menge von Instanzbezeichnern gilt, kann sich beispielsweise die Konzeptzugehörigkeit der jeweiligen Instanzen im Laufe der Analyse kontextabhängig durch Spezialisierung

¹⁶Erst die Übertragung der Ergebnisse des Konzeptlernens in die Wissensbasen für das Upper-Model, die Medizin oder die Informationstechnologie führt zu einer inkrementellen Erweiterung der terminologischen Beschreibungen während der Analyse (Hahn & Schnattinger, 1997). Dies wird hier nicht weiterverfolgt.

ändern. Im Anschluß an die Einführung des Instanzsymbols in das Diskursuniversum als Anker für die Referenz erfolgt die eigentliche typisierende Instantiierung in einem *Initialkontext*. Dazu wird ein direkter Subkontext unter NEWTEXT generiert und die aus dem Lexikoneintrag *lex* im Lexikon $\mathcal{LE}$ hervorgehende Information über die Konzeptzugehörigkeit zur Instantiierung verwendet. Die Definition 4.4 beinhaltet eine genauere Beschreibung der Eigenschaften eines Initialkontexts.

Definition 4.4 *Kontexte $h \in \mathcal{H}$, in denen die konzeptuellen Korrelate $lex.con$ (mit $con \in 2^{\mathcal{F}}$) eines lexikalischen Eintrags $lex \in \mathcal{LE}$ für eine Wortoberfläche instantiiert werden, sind Initialkontexte. Per Konvention tragen diese Initialkontexte in ihrem eindeutigen Bezeichner den Teilstring LEXHYPO, wenn zu einer Wortoberfläche nur eine Wortklasse oder nur ein Konzeptidentifizier bekannt ist, ansonsten wird der Teilstring LEXAMBIGHYPO verwendet. Initialkontexte stehen in einer direct-subcontext-of-Beziehung zur initialen Textwissenbasis NEWTEXT. Syntaktisch gesehen überdecken die in ihnen interpretierten Strukturen genau einen Knoten im Dependenzgraphen. Konzeptuell betrachtet fügen die Initialkontexte den terminologischen und assertionalen Axiomen der Ausgangswissensbasis genau eine Assertion in Form einer Konzeptinstantiierung hinzu.*

Auf der Wortebene kommt die Ambiguitätsproblematik zweifach ins Spiel: Eine Wortoberfläche kann auf mehrere Lexikoneinträge verweisen und zu einem Lexikoneintrag können mehrere Konzeptidentifizier vorliegen. Aufgrund des unterschiedlichen syntaktischen Bindungsverhaltens im Satz, das sich durch die an Dependenzrelation geknüpfte semantische Interpretation begründet, ist für jede Wortarthypothese ein separater Initialkontext vorzuhalten (vgl. Kapitel 4.1). Sind mit einem Lexikoneintrag überdies noch mehrere Konzeptidentifizier assoziiert, dann muß ebenfalls für jeden dieser Identifizier ein eigener Initialkontext generiert werden. Konkret bedeutet dies, daß bei einer mit i Lexikoneinträgen verbundenen Wortoberfläche, bei der jeder Lexikoneintrag mit j Konzeptidentifiern verbunden ist, genau i mal j Initialkontexte zu generieren sind. Demgemäß entsteht mit der Instantiierung von Wortoberflächen unter Umständen eine große Anzahl von lokalen Ambiguitäten. Die meisten hiervon werden allerdings bei der Integration in den Satz nicht weiterverfolgt, weil der syntaktische Kontext ebenso wie die semantischen Constraints die weitere Analyse restringieren. Eine genauere, beispielorientierte Beschreibung der lokalen Ebene von Ambiguitäten findet sich in nachfolgenden Kap. 4.3.4.

Die beschriebene Erzeugung der Initialkontexte fällt mit der Instantiierung der Wortaktoren auf der Seite des Parsers zusammen. Zum Instantiierungszeitpunkt wird ein Wortaktor von einem Phrasenaktor umschlossen. Obwohl die Instantiierung von Diskursreferenten in der Textwissensbasis von den Wortaktoren initiiert wird, verwalten generell die Phrasenaktoren die Referenzen auf Kontexte in der Textwissensbasis (siehe Definition 4.5). Im Falle von Initialkontexten besteht die aktuelle Interpretation zunächst nur aus den Instantiierungen der Diskursreferenten. Die Phrasenaktoren wiederum befinden sich innerhalb eines Containeraktors. Treten Ambiguitäten auf der Ebene der Wortklassenzugehörigkeit auf, dann obliegt deren Verwaltung den Containeraktoren, wie in Kapitel 3.2.1 aufgezeigt wurde.

Definition 4.5 *Bezeichne P die Menge aller Phrasenaktoren, dann existiert zu jedem Phrasenaktor $p \in P$, der Teilberechnungsergebnisse auf syntaktischer Seite kapselt, eine (möglicherweise leere) Menge $cont \in 2^{\mathcal{H}}$ von Kontexten im Kontextgraphen, notiert als $p.cont$, welche die semantische Interpretation des im Phrasencontainer befindlichen Dependenzgraphen repräsentieren.*

Auf der Grundlage des im Kapitel 4.1.1 eingeführten Begriffs des “minimalen interpretierbaren Subgraphen” ist es nun möglich anzugeben, wann die Generierung neuer Interpretationskontexte im Lauf der inkrementellen Analyse eines Eingabetexts notwendig wird. Seien $cont_1$ und $cont_2$ Mengen von Interpretationskontexten aus $2^{\mathcal{H}}$ und p_1 und p_2 Phrasenaktoren aus P . Wenn ein Wortaktor in p_1 mit einem Wortaktor in p_2 eine Dependenzrelation eingeht (d.h. *alle* an die Valenz des Kopfes geknüpften Wohlgeformtheitsbedingungen sind erfüllt) *und* dadurch ein vollständiger minimaler interpretierbarer Subgraph entsteht, dann müssen die aus der semantischen Interpretation resultierenden Ergebnisse als neue Hypothesen in einem Kontext abgelegt werden. Zur semantischen Interpretation müssen alle Tupel von Kontexten, die sich aus dem Cartesischen Produkt $cont_1 \times cont_2$ ergeben, herangezogen werden, da jedes dieser Tupel für eine alternative Lesart steht.¹⁷ Daraus ergibt sich für die Interaktion zwischen Syntax und semantischer Interpretation folgendes Protokoll:

¹⁷Da die semantische Interpretationsfunktion den Konzepttyp der in der Textwissensbasis in die Interpretation involvierten Instanzen verwendet, muß auch die (in Kapitel 4.1 eingeführte) Funktion *type* kontextabhängig definiert werden. D.h., *type* liefert den Konzepttyp zu einem Instanzsymbol in Abhängigkeit zu einem Kontext.

1. Wenn durch eine angefragte Dependenzrelation ein semantisch interpretierbarer Subgraph vollständig wird, dann extrahiere die Kontextmengen $p_1.cont_1$ und $p_2.cont_2$ der beiden beteiligten Phrasenaktoren p_1 und p_2 .
2. Instantiiere für jedes Tupel in $cont_1 \times cont_2$ das zum minimalen interpretierbaren Subgraphen gehörige semantische Interpretationsschema, indem für jede der beiden involvierten Instanzen i_1 und i_2 in Abhängigkeit vom Kontext der Konzepttyp mit der Funktion *type* ermittelt wird.
3. Seien h und h' Kontexte und sei $\langle h, h' \rangle$ ein Tupel in $cont_1 \times cont_2$, für das das semantische Interpretationsschema instantiiert wird. Sei n die Anzahl der Relationspfade, die die semantische Interpretation in dem durch h und h' gegebenen Interpretationskontext berechnet (vgl. Kap. 4.1.1.1). Generiere nun n neue Kontexte $h_1, h_2, \dots, h_n$, so daß für jedes $i \in \{1, 2, \dots, n\}$ gilt: *direct-subcontext-of*(h_i, h) $\wedge$ *direct-subcontext-of*(h_i, h').
4. Sei REL_i ein durch die semantische Interpretation berechneter Relationspfad aus der Menge der n berechneten Relationspfade. Da jeder dieser Relationspfade eine unterschiedliche Lesart beinhaltet, kapselt ein Kontext die formale Repräsentation dieser interpretatorischen Alternative in der Form $(i_1 REL_i i_2)_{h_i}$.
5. Ermittle im Anschluß an die Instantiierung der semantischen Interpretationsfunktion für alle Tupel in $cont_1 \times cont_2$ die Menge $cont_3$ von Interpretationskontexten des aktuell verarbeiteten semantisch interpretierbaren Subgraphen, die die neu berechneten Ergebnisse beinhalten. Dazu sind im erweiterten Konzeptgraphen zur Menge $cont_1 \cup cont_2$ diejenigen Subkontexte zu berechnen, die gleichzeitig *terminale* Kontexte und nicht bereits in $cont_1 \cup cont_2$ enthalten sind.¹⁸ Diese definieren die Menge $cont_3$.
6. Da das Einziehen der angefragten Dependenzrelation zur Instantiierung eines neuen Phrasenaktors p_3 führt, erfolgt die Assoziierung der Menge der Interpretationskontexte $cont_3$ mit dem Phrasenaktor in der Form $p_3.cont_3$. Durch diese Form der Verknüpfung von semantischer Interpretation mit den Berechnungen des Parsers bleiben

¹⁸Bedingt durch konzeptuelle Interpretationen, deren Assertionen ebenfalls in separaten Kontexten gekapselt sind, können weitere Kontexte unterhalb von $cont_1$ und $cont_2$ entstehen.

die Mehrdeutigkeiten in der Textwissenbasis dem Parser verborgen, was in Definition 4.6 als Opaheits-Prinzip beschrieben ist.

In Abb. 4.11 findet sich ein Beispiel für das aufgezeigte Verfahren. Die Menge an Interpretationskontexten für zwei Phrasenaktoren, p_1 und p_2 , sei gegeben durch $p_1.cont_1 := \{h_1, h_2\}$ und $p_2.cont_2 := \{h_3, h_4, h_5\}$. Die Instantiierung der semantischen Interpretationsfunktion für jedes Tupel aus $cont_1 \times cont_2$ ergäbe für $\langle h_1, h_3 \rangle$ zwei Relationspfade, für $\langle h_2, h_5 \rangle$ einen Relationspfad und keinen Relationspfad in allen anderen Fällen. Jede der berechneten Interpretationen kommt nun in einen separaten Kontext, welcher als Subkontext des dazugehörigen Tupels an Interpretationskontexten angelegt wird. Die Vereinigungsmenge aller auf diese Weise generierten Kontexte ergibt $p_3.cont_3 := \{h_6, h_7, h_8\}$.

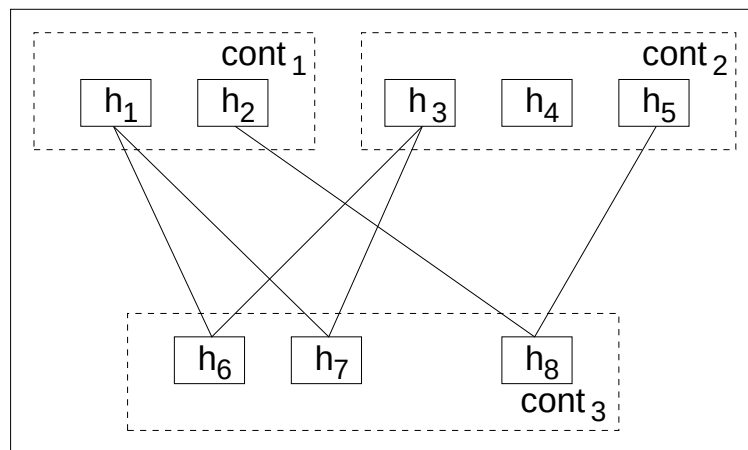


Abbildung 4.11: Kontexte und semantische Interpretation

Definition 4.6 Opaheits-Prinzip: Die durch Polysemie oder durch semantische Interpretation bedingten Ambiguitäten werden auf der Seite der Textwissenbasis durch Mengen an Kontexten verwaltet, die disjunkte Hypothesen über den Textinhalt repräsentieren. Sie bleiben somit für den Parser opaque und interferieren nicht mit der syntaktischen Analyse. Jeder Kontext gehört zu höchstens einem Phrasenaktor. Einem Phrasenaktor können indessen mehrere Kontexte zugeordnet sein, die semantische Interpretationsalternativen kapseln. Dabei gilt, daß die mit Phrasenaktoren verbundenen Kontextmengen kein gemeinsames Kontextelement besitzen: $\forall p_1, p_2 \in P : p_1.cont_1 \cap p_2.cont_2 = \emptyset$ mit $p_1 \neq p_2$ und $cont_1, cont_2 \in 2^{\mathcal{H}}$.

The screenshot shows a 'Lexicon Browser' window with the title bar path: /usr/local/coling/data/lexica/upper-model/verbs-surf.lex. The window is divided into several panes:

- lexicon entry keys:** A list of verb forms including 'entwickelt', 'erfassen', 'erfaßt', 'erfaßte', 'erfolgt', 'erfordern', 'ergeben', 'ergibt', 'erhalten', 'erhofft', 'erhält', 'erkennen', 'erkennt', 'erleichtern', 'ermittelt', 'ermöglichen', and 'ermöglicht'. 'entwickelt' is selected.
- ambiguities:** A list showing the base form 'entwickeln' for the selected entry.
- surface:** Displays the selected form 'entwickelt'.
- word class:** Displays 'VerbFinite'.
- lexeme:** Displays 'entwickeln'.
- individual features:** A detailed feature structure for the selected form:


```

      {agr: {[num: pl
              pers: 2]
            [num: sg
              pers: 3]}
        mood: ind
        tense: pres}
      
```

Abbildung 4.12: Beispiel für ambige Wortklassenzugehörigkeit eines Lexikoneintrags

Innerhalb des Kontextgraphen, der sich im Laufe der inkrementellen Textanalyse aufbaut, finden sich die inhaltlichen Interpretationen zu Teilberechnungen wie auch das Ergebnis des Gesamttexts grundsätzlich in den *terminalen Kontexten*. Über die Assoziation mit Phrasenaktoren kann darüber hinaus prinzipiell zu jedem Zeitpunkt während der inkrementellen Analyse auf die in den terminalen Kontexten befindlichen konzeptuellen Repräsentationen zugegriffen werden. Sei $w_1, w_2, \dots, w_n$ ein Folge von Wörtern innerhalb eines Texts. Nach der semantischen Interpretation jedes Wortinkrements w_i mit $i \in \{1, 2, \dots, n\}$ existiert eine Menge terminaler Kontexte, die die Interpretation des Eingabetexts einschließlich aller alternativen Lesarten beinhalten. Auf der Satzebene schließlich finden sich mit der Konsumierung des Punktes, d.h. dem Abschluß der Einzelsatzanalyse, alle (Teil-)Interpretationen in den zum Einzelsatz gehörenden terminalen Kontexten des Kontextgraphen.

Die nun nachfolgenden Unterkapitel 4.3.5 und 4.3.7 geben eine beispielorientierte Darstellung der Interaktion der Phrasenaktoren mit den Kontexten der Textwissensbasis.

4.3.4 Behandlung lexikalischer Ambiguitäten

Lexikalische Ambiguitäten kommen auf zwei verschiedenen Ebenen vor. Erstens kann eine Wortoberfläche – rein lokal betrachtet – *unterschiedlichen Wortklassen* angehören. Auf der syntaktischen Seite führt dies dazu, daß für jede dieser Wortklassen eine Wortaktoreninstanz erzeugt wird. Dies ist notwendig, weil sich jede Wortklasse durch unterschiedliche Valenzdefinitionen syntaktisch anders verhält, wodurch unterschiedliche syntaktische

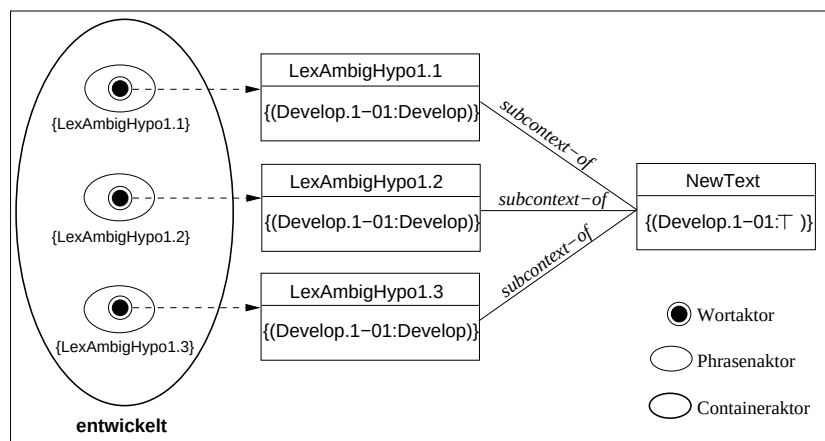


Abbildung 4.13: Lexikalische Ambiguität I (Part-of-Speech): Aktoren und Initialkontexte

Strukturen mit jeweils eigenen semantischen Interpretationen berechnet werden. Darüber hinaus bestimmt die Wortklasse auch die Merkmalsstruktur eines Wortes.

In Abb. 4.12 ist der Lexikoneintrag zu “entwickelt” dargestellt, so wie ihn der *Lexikon Browser*, das grafische interaktive Werkzeug von SYNDIKATE zur Lexikonbearbeitung, darstellt. Auf der Ebene der Wortklassenzugehörigkeit finden sich drei Alternativen. Neben den Wortklassen *VerbFinite* und *VerbPartPassive* findet sich die Wortklasse *VerbPartPerfect* für die partizipiale Form in analytischen Tempuskonstruktionen. Bei der in Abb. 4.12 angezeigten Wortform handelt es sich um das finite Verb. Das Browserfeld **individual features** beinhaltet die Merkmalsstruktur des Eintrags sowie Angaben über den Tempus und den Modus. Daneben enthält das Feld für **lexeme** einen Verweis auf die Grundform im Lexemlexikon, das einen Teil der Worteigenschaften definiert. Im Unterfenster **ambiguities** finden sich außer dem aktuell aktiven Eintrag die beiden anderen Einträge für *VerbPartPassive* und *VerbPartPerfect*. Diese drei Einträge weisen jedoch entsprechend ihrer voneinander abweichenden Wortklassenzugehörigkeit unterschiedliche Merkmalsstrukturen und einen unterschiedlichen Bestand an Valenzen auf.

An die sich daraus ergebenden Bindungseigenschaften im Satzkontext geknüpft ist eine divergierende semantische Interpretation der jeweiligen Abhängigkeitsrelationen (vgl. Kap. 4.1). Daraus folgt, daß für jede lexikalische Oberfläche, die mit mehr als einem lexikalischen Eintrag $lex \in \mathcal{LE}$ (d.h. mit mehr als einer Wortklasse) verbunden ist, für jede dieser Varianten in der Textwissensbasis eine eigene Interpretation vorgehalten werden muß genau dann, wenn für lex mindestens ein Konzeptidentifizierer bekannt ist. Jede dieser Interpretati-

onsvarianten befindet sich dann in einem separaten Initialkontext. Dazu wird zunächst ein eindeutiger Identifierstring für die zu erzeugende Instanz generiert. In den Ausgangskontext der Textwissenbasis NEWTEXT kommt dann eine Assertion, die das Instanzsymbol ungetypt in das Diskursuniversum einführt. Anschließend wird ein Initialkontext für jede Wortklassenambiguität als direkter Subkontext zur initialen Textwissenbasis NEWTEXT definiert, in die die Instanz mit dem im Lexikon enthaltenen Konzepttyp eingestellt wird.

Abb. 4.13 enthält den hier aufgezeigten Zusammenhang für die Wortoberfläche “entwickelt”. Entsprechend der in Abb. 4.12 dargestellten Mehrdeutigkeit des Lexikoneintrags auf der Ebene der Wortklassenzugehörigkeit sind auf der syntaktischen Seite zunächst drei Instanzen eines Wortaktors anzutreffen. Jeder dieser Wortaktoren repräsentiert eine Wortklasse. Die Wortaktoren wiederum sind in Phrasenaktoren eingebettet, der Containeraktor hingegen kapselt die Part-of-Speech-Ambiguität (vgl. Abb. 4.13 linke Seite). Jedem dieser Wortaktoren ist nun ein Kontext zugeordnet, der die unterschiedliche Bindungsfähigkeit mit ihren Implikationen für die semantische Interpretation kapselt. Diese Initialkontexte sind Subkontexte der anfänglichen Textwissenbasis und haben einen eindeutigen Identifier (LEXAMBIGHYPO1.1, LEXAMBIGHYPO1.2 und LEXAMBIGHYPO1.3). Das generierte Instanzsymbol DEVELOP.1-01 wird nun entsprechend der im Lexikon verzeichneten Konzeptzugehörigkeit der Oberflächenform “entwickelt” zum Konzepttyp DEVELOP in allen Initialkontexten instantiiert. Zugleich erfolgt die Einführung des Diskursreferenten durch ein ungetyptes Instanzsymbol in NEW-TEXT, der initialen Textwissenbasis.

Die zweite Form der lexikalischen Ambiguität betrifft die semantische Mehrdeutigkeit einer Wortoberfläche bei identischer Wortklasse (Polysemie). Systemseitig wird dies dadurch kodiert, daß eine lexikalische Oberfläche im Lexemlexikon mit mehr als einem Konzeptidentifier assoziiert ist. Enthält demzufolge die zu einem Lexikoneintrag $lex \in \mathcal{LE}$ gehörende Menge $con \in 2^{\mathcal{F}}$ mehr als ein Element, dann müssen im lokalen Kontext der inkrementellen Analyse alle diese Bedeutungsvarianten in einem separaten Kontext instantiiert werden.

Im Gegensatz zu den Ambiguitäten bei Wortklassenzugehörigkeit sind einem Wortaktor, dessen Existenz fest mit den Wortarten verknüpft ist, mehrere Initialkontexte zugeordnet, um genau die *semantische* Mehrdeutigkeit zu repräsentieren. Die Attribuierung von mehr als einem Kontext zu einem Phrasenaktor, um die Bedeutungsverschiedenheit von Wörtern (und später auch von Phrasen und ganzen Sätzen) zu repräsentieren, garantiert eine strikte Trennung von syntaktisch strukturellen Mehrdeutigkeiten einerseits und semantisch konzeptuellen Mehrdeutigkeiten andererseits. Darüber hinaus bleiben syntax-

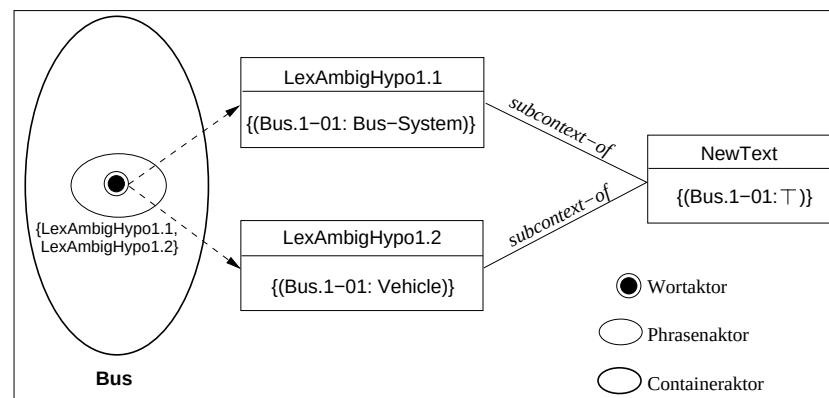


Abbildung 4.14: Lexikalische Ambiguität II (Polysemie): Aktoren und Initialkontexte

unabhängige Interpretationsvarianten eines Textes auf der syntaktischen Ebene opaque und interferieren nicht mit der eigentlichen syntaktischen Analyse.

In Abb. 4.14 kommt genau dieser Zusammenhang zur Darstellung. Gegeben sei das Wort *“Bus”*. Auf der Ebene der Wortklassenzugehörigkeit handelt es sich bei dieser Wortoberfläche eindeutig um ein Nomen. Für die Bedeutungsrepräsentation nehmen wir zwei Lesarten an. Zunächst einen *“Bus”* im Sinne eines Fahrzeug zum Personentransport, zum anderen *“Bus”* als Hardwarekomponente, wie sie in der IT häufig anzutreffen ist. Damit hängt die semantische Interpretation im Satzkontext einzig von den unterschiedlichen Bedeutungen und nicht von der Syntax ab, wie man sich an den Phrasen *“der Bus des Rechners”* versus *“der Bus des Fuhrunternehmers”* leicht vergegenwärtigen kann. Bei identischer syntaktischer Struktur kann jeweils nur eine Bedeutungsvariante des Nomens semantisch richtig interpretiert werden. Im Lexikon verweist der Eintrag für die Wortoberfläche entsprechend auf zwei Konzeptidentifizier, nämlich *VEHICLE* und *BUS-SYSTEM*. Für beide Lesarten erzeugt das System nun dieselbe, eindeutige Individuenkonstante, *BUS.1-01*. Nach der ungetypten Einführung der Individuenkonstanten in den Kontext *NEWTEXT* (vgl. Abb. 4.14 rechte Seite) erfolgt jeweils die Instantiierung mit dem jeweiligen Konzepttypen in einem separaten Kontext (*LEXAMBIHYP01.1* bzw. *LEXAMBIHYP01.2*). Der Kontextmechanismus ermöglicht eine unterschiedliche (kontextabhängige) Interpretation ein und derselben Individuenkonstante. Die beiden, die unterschiedlichen Lesarten trennenden Kontexte sind indessen demselben Phrasenaktor zugeordnet (vgl. Abb. 4.14 linke Seite). Zwar initialisieren die Wortaktoren die Generierung der Initialkontexte, die eigentliche Ver-

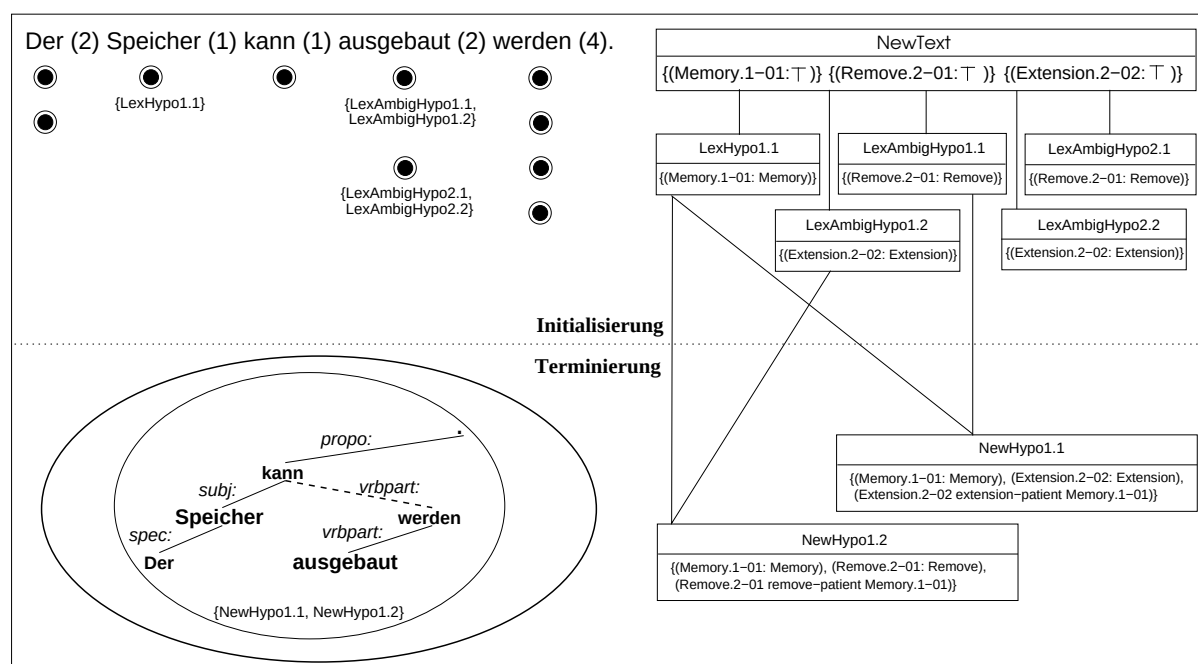


Abbildung 4.15: Repräsentation semantischer Interpretationsalternativen durch Kontexte

waltung der Kontexte auf der Weltwissenseite, die Interpretationszustände während der Textanalyse widerspiegeln, obliegt indessen den Phrasenaktoren.

Da die lexikalischen Ambiguitäten auf der Ebene der Wortklassenzugehörigkeit gehandhabt wird, indem Wortaktoren instantiiert werden, und die lexikalischen Ambiguitäten auf der semantischen Ebene der Wortbedeutung gehandhabt werden, indem Konzepte in Kontexten instantiiert werden, sind beide Ambiguitätsdimensionen bei Wortoberflächen durch unterschiedliche Repräsentationsschichten (Syntax und Semantik) klar getrennt.

4.3.5 Semantische Ambiguitäten bei der Satzanalyse

Im Anschluß an die Instantiierung von Diskursentitäten in den Initialkontexten erfolgen im Verlauf der inkrementellen Analyse eines Eingabetexts semantische Interpretationen, sobald zwei Diskursentitäten über einen interpretierbaren Subgraphen zusammengebracht werden. Hierbei verarbeitet der Kontextmechanismus die möglicherweise auftretenden Ambiguitäten in Form mehrerer Relationspfade bzw. in Form der Weiterverarbeitung von Polysemien.

In Abb. 4.15 kommt die Interaktion zwischen dem Kontextmechanismus auf der Seite

der Textwissensbasis und den Phrasen- und Containeraktoren auf Seiten des Parsers zur Darstellung. Zur Vereinfachung sind abweichend vom eigentlich inkrementellen Analyseverfahren die Instantiierung von Wortaktoren sowie der dazugehörigen Initialkontexte bereits für den gesamten Beispielsatz vorgenommen. In der linken Hälfte der Abbildung finden sich die syntaktischen Objekte, während die konzeptuellen auf der rechten platziert sind. Für die Syntax kommt als weitere Vereinfachung hinzu, daß aus Platzgründen nur die Instanzen des Wortaktoren, nicht aber die sie jeweils umschließenden Phrasen- und Containeraktoren in der Abbildung enthalten sind. Ambiguitäten bezüglich der Wortklassenzugehörigkeit einer lexikalischen Oberfläche spiegeln sich in der Anzahl der instantiierten Wortaktoren wider. Beispielsweise existiert „*der*“ im Lexikon einmal als Relativpronomen und einmal als definiter Artikel. Im kleinen Beispielsatz gibt es nur zwei Inhaltswörter, nämlich „*Speicher*“ und „*ausgebaut*“. Interessant ist vor allem „*ausgebaut*“, da dieses Wort sowohl auf der Ebene der Wortklassen (*VerbPartPassive* und *VerbPartPerfect*) als auch auf der Ebene der Bedeutung mehrdeutig ist. Syntaktisch wird die Mehrdeutigkeit durch die Anzahl der Wortaktoren repräsentiert, in der Textwissensbasis durch die Kreierung von Instanzen mit unterschiedlichem Konzepttyp in einem eigenen Kontext während der Initialisierungsphase. Das Wort „*ausbauen*“ ist polysem, denn einerseits kann es im Sinne von „*erweitern*“ andererseits im Sinne von „*entfernen*“ interpretiert werden. Daher gilt z. B. (REMOVE.2-01 : REMOVE)_{LexAmbigHypo1.1} und (EXTENSION.2-02 : EXTENSION)_{LexAmbigHypo1.2}.

10. Der Speicher kann ausgebaut werden.

Mehrdeutigkeiten semantischer Herkunft sind für den Parser unsichtbar. Deren Verwaltung findet allein auf der Ebene der Kontexte in der Textwissensbasis statt. Entsprechend kennt der Phrasenaktor nur eine Menge, die die Referenzen auf die zum Phrasenaktor gehörigen Interpretationskontexte speichert. Wieviele dieser Kontextobjekte in der Menge enthalten sind, ist dem Phrasenaktor unbekannt. Entsprechend der Doppeldeutigkeit von „*ausgebaut*“ sind mit jedem der Phrasenaktoren für eine Wortartzugehörigkeit der Oberflächenform zwei Initialkontexte verbunden, d.h. LEXAMBIGHYPO1.1 und LEXAMBIGHYPO1.2 bzw. LEXAMBIGHYPO2.1 und LEXAMBIGHYPO2.2 (siehe Abb. 4.15 linke Seite oben). Im Gegensatz dazu verfügt der Phrasenaktor für „*Speicher*“ nur über einen Initialkontext (LEXHYPO1.1), da weder eine Ambiguität bezüglich der Wortklassenzugehörigkeit noch eine Polysemie vorliegt.¹⁹

¹⁹Dies ist natürlich eine Vereinfachung aus Darstellungsgründen, da ein Speicher bsp. ebenfalls einen

Im Laufe der inkrementellen Analyse ist der nächste Zeitpunkt einer Interaktion von Parser mit den Interpretationskontexten in der Textwissenbasis, die hier alle noch Initialkontexte sind, gegeben, sobald die Dependenzrelation *vrupart* zwischen *“kann”* und *“werden”* berechnet wird. Durch die Prüfung der Wohlgeformtheitsbedingungen für die *vrupart*-Valenz wird ein minimaler semantisch interpretierbarer Subgraph vollständig, d.h., die semantische Interpretation wird ausgelöst. Auf der Seite der Kontexte wird dazu eine Anfrage zwischen den zu relationierenden Diskursreferenten in der Textwissenbasis (i.e. MEMORY.1-01 und REMOVE.2-01 bzw. EXTENSION.2-02) in Abhängigkeit zu den Interpretationskontexten generiert. Gegeben sind diese Interpretationskontexte durch den Phrasenaktor mit dem anfragenden Modifikator und dem(n) Phrasenaktor(en) des linken syntaktischen Kontexts. Im vorliegenden Beispiel wären dies die Kontexte LEXAMBIGHYPO1.1 und LEXAMBIGHYPO1.2, die mit dem Phrasenaktor für die passivische partizipiale Lesart der Teilphrase *“werden ausgebaut”* als Passivinfinitiv verbunden sind, und der Kontext LEXHYPO1.1 als inhaltliche Repräsentation der Teilphrase *“der Speicher kann”*.²⁰ Gemäß dem beschriebenen Verfahren zum Management der Interpretationskontexte findet nun im Zuge der ausgelösten semantischen Interpretation eine Prüfung der konzeptuellen Relationierbarkeit zwischen allen aus den beiden Kontextmengen zu bildenden Kontexttupeln statt. Im Beispiel beschränkt sich dies auf die zweimalige Instantiierung des Interpretationsschemas für Modal- und Auxiliärverben. Das durch die Interpretationsfunktion ermittelte Ergebnis führt, da beide konzeptuellen Lesarten zulässig sind, im Resultat zur Relationierung der involvierten Instanzen über unterschiedliche Relationen in separaten Interpretationskontexten:

(EXTENSION.2-02 EXTENSION-PATIENT MEMORY.1-01)_{NewHypo1.1}

(REMOVE.2-01 REMOVE-PATIENT MEMORY.1-01)_{NewHypo2.1}

Da es sich bei den neu aufgebauten Interpretationskontexten um Interpretationsalternativen für eine einzige syntaktische Struktur handelt, werden entsprechend dem Prinzip, semantische Ambiguitäten nicht mit syntaktischen interferieren zu lassen, beide Kontexte, NEWHYPO1.1 und NEWHYPO2.1, dem durch die neue Dependenzrelation instantiierten

Raum in einem Gebäude bezeichnet. Die Einbeziehung dieser Lesart ergäbe – ohne den textuellen Kontext – eine größere Anzahl an Interpretationskontexten. Die Beschreibung des grundlegenden Verfahrens wird durch diese Vereinfachung indessen nicht affiziert.

²⁰Da syntaktisch nur die passive Lesart erzeugt wird, finden mit den anderen Initialkontexten keine weiteren semantischen Interpretationen statt.

Phrasenaktor zugeordnet (siehe Abb. 4.15 linke Seite unten). Die Ergebnisse der semantischen Interpretation für einen einzelnen Satz befinden sich mit der Konsumierung des Punkts in den terminalen Kontexten, d.h. in NEWHYP01.1 und NEW-HYP01.2.

Zu Ende dieses kleinen Kapitels sei angefügt, daß die Ergebnisse der konzeptuellen Interpretation analog zu denen der semantischen Interpretation behandelt werden. Auch die durch die konzeptuelle Interpretation berechneten Assertionen kommen in separate Kontexte im Kontextgraphen. Weil die konzeptuelle Interpretation datengetrieben ist (durch Hinzufügen von Assertionen), erfolgen sie im Anschluß an die semantische Interpretation und erweitern den Kontextgraphen um neue terminale Kontexte, die dann die aktuellen Interpretationsergebnisse halten.

4.3.6 Interpretationskontexte und Disambiguierung

Auch wenn die semantische Interpretation einen oder mehrere gültige konzeptuelle Pfade ermittelt, kann das Hinzufügen des entsprechenden assertionalen Axioms in einen Interpretationskontext zu einer inkonsistenten A-Box führen. Diese Inkonsistenz ist dann ein Indikator dafür, daß die in diesem Interpretationskontext gehaltene Lesart kein gültiges Interpretationsergebnis darstellt und daher von der weiteren Verarbeitung auszuschließen ist.

Um zu verdeutlichen, wie ein inkonsistenter Interpretationskontext entsteht, führt Satz 11 ein kurzes Beispiel ein.

11. Die Aufrüstung des Computers von OS1.2 auf OS2.0

Der Interpretation liegt eine kleine, sehr einfach gehaltene Wissensbasis zugrunde, die aus einem Konzept UPGRADE (für die Wortoberfläche *“Aufrüstung”*) besteht, das eine definitonische single-value Rolle UPGRADE-PATIENT besitzt, die also nur einen Füller haben kann. Darüber hinaus gibt es zwei Relationen am Konzept UPGRADE, die die obere und untere Grenze einer UPGRADE-Handlung bestimmen (LOWER-BOUND und UPPER-BOUND). Die Typbeschränkung der UPGRADE-PATIENT-Relation sei PRODUCT. Das Konzept PRODUCT wiederum subsumiert die Konzepte COMPUTER-SYSTEM und OPERATING-SYSTEM (der Konzepttyp für *“OS1.2”* bzw. *“OS2.0”*).

Im folgenden betrachte ich nur die beiden Interpretationsschritte, die durch die minimalen Subgraphen (*“Aufrüstung”* – *genitive-attribute* – *“Computers”*) und (*“Aufrüstung”* – *ppcomp/ppattribute* – *“von”* – *pobject* – *“OS1.2”*) gegeben sind.

Im ersten Interpretationsschritt für das Genitivattribut wird versucht, UPGRADE und COMPUTER-SYSTEM miteinander zu relationieren. Der semantische Interpretationsprozeß terminiert mit der Relation UPGRADE-PATIENT, die entsprechende Assertion wird in einen neuen Kontext eingestellt.

Wenn nun die Präpositionalphrase mit *“von”* versucht, sich an den syntaktischen Kopf *“Aufrüstung”* anzubinden, dann gibt es eine syntaktische Bindungsambiguität. Entweder die Präpositionalphrase ist als Präpositionalobjekt zu interpretieren oder aber als Präpositionalattribut. Die Mehrdeutigkeit rührt daher, daß *“Aufrüstung”* eine Verbnominalisierung ist. Daher kann eine thematische Rolle vom Typ AGENT oder PATIENT als Präpositionalobjekt durch eine *“von”*-Präpositionalphrase eingeführt werden. Der Attributfall schließt hingegen (über die entsprechende Restriktionsmenge untersagter Relationen, die an der Valenz definiert sind) genau diese Interpretationen als AGENT oder PATIENT aus.

Was passiert nun im Attributfall? Da die konzeptuelle Relation LOWER-BOUND in der Restriktionsmenge der Präposition *“von”* enthalten ist, kann die entsprechende Anfrage, ob UPGRADE und OPERATING-SYSTEM relationierbar sind, positiv mit Rückgabe der Relation LOWER-BOUND abgearbeitet werden. Dementsprechend wird eine Assertion in einen neuen Kontext eingefügt. Die andere Lesart als Präpositionalobjekt, die versucht UPGRADE und OPERATING-SYSTEM als AGENT oder PATIENT zu relationieren, gibt auch ein positives Ergebnis zurück: UPGRADE-PATIENT.

Der entscheidende Unterschied zum Attributfall besteht darin, daß das zusätzliche assertionale Axiom die Relation UPGRADE-PATIENT zum zweiten Mal belegt. Da die Relation indessen nur einen Füller zuläßt, führt diese Assertion zu einem inkonsistenten Interpretationskontext. Aus diesem Grund ist die Interpretation unzulässig und wird im folgenden nicht weiterverfolgt. Als Konsequenz ist die in Verbindung mit der semantischen Interpretation stehende Dependenzrelation *ppcomp* ebenfalls nicht zulässig, denn die Anfrage resultiert in keinem validen Interpretationskontext.

Die fortlaufende Prüfung, ob Assertionen, die sich durch die semantische Interpretation ergeben, auch in einer konsistenten A-Box in den jeweiligen Interpretationskontexten resultieren, dient der Disambiguierung. Interpretationen, für die es kein Modell gibt, sind vom weiteren Fortgang der Textanalyse auszuschließen.

4.3.7 Behandlung syntaktischer Ambiguitäten

Strukturelle Ambiguitäten werden in der Textwissensbasis dadurch reflektiert, daß das unterschiedliche syntaktische Bindungsverhalten unterschiedliche semantische Interpretationen triggert. D.h., die semantischen Interpretationsschemata erhalten in Abhängigkeit vom syntaktischen Kontext andere Instantiierungsparameter sowohl hinsichtlich der zu relationierenden Entitäten in der Textwissensbasis als auch hinsichtlich der die Interpretation steuernden statischen Constraints.

Ein Beispiel zur Veranschaulichung der Behandlung struktureller Ambiguitäten in der Textwissensbasis mithilfe von Kontexten findet sich in Abb. 4.16. Die Verarbeitung des Eingabesatzes *“Geliefert wird der Computer von Compaq”* unter Interaktion von Syntax mit der Textwissensbasis wird in drei Schritten skizziert. Zunächst kommen die Objekte beider Ebenen zur Instantiierungszeit zur Darstellung (Abb. 4.16 oben). Anschließend wird das Ergebnis nach der Analyse des ersten vollständigen semantisch interpretierbaren Subgraphen wiedergegeben (Abb. 4.16 Mitte). Die abschließende inhaltliche Interpretation des Eingabesatzes findet sich in den terminalen Phrasenaktoren des Parsers bzw. den terminalen Kontexten der Textwissensbasis (Abb. 4.16, unten links bzw. rechts). Zur Vereinfachung der Präsentation wird in der Initialisierungsphase der Aspekt der Inkrementalität wieder vernachlässigt. Auf syntaktischer Seite wird für jede mögliche Wortart zu einer Wortoberfläche jeweils ein Wortaktor instantiiert. Jeder Wortaktor ist – was in der Abbildung nicht dargestellt wird – zur Instantiierungszeit von einem Phrasen- und einem Containeraktor umschlossen. Entsprechend dem Mechanismus zur Verbindung von Wortaktoren mit Diskursreferenten erfolgt auf Seiten der Textwissensbasis die Instantiierung der mit den Lexikoneinträgen assoziierten Konzeptidentifizier in je separaten Initialkontexten. Dabei liegen keine Polysemien vor. Diese Initialkontexte wiederum sind je einem Phrasenaktor zugeordnet.

Mit der Konsumierung des Worts *“Computer”* entsteht ein semantisch interpretierbarer Subgraph, der sich aus vier lexikalischen Knoten zusammensetzt: *“Geliefert wird der Computer”*. Durch die Instantiierung des semantischen Interpretationsschemas für Passivkonstruktionen erfolgt in der Textwissensbasis die Berechnung einer konzeptuellen Relationierung (DELIVER-PATIENT) zwischen den beiden involvierten Diskursentitäten DELIVERY.1-01 und COMPUTER.2-01, welche in einem neu generierten Kontext (NEWHYPO1.1) eingezogen wird (Abb. 4.16, Mitte).

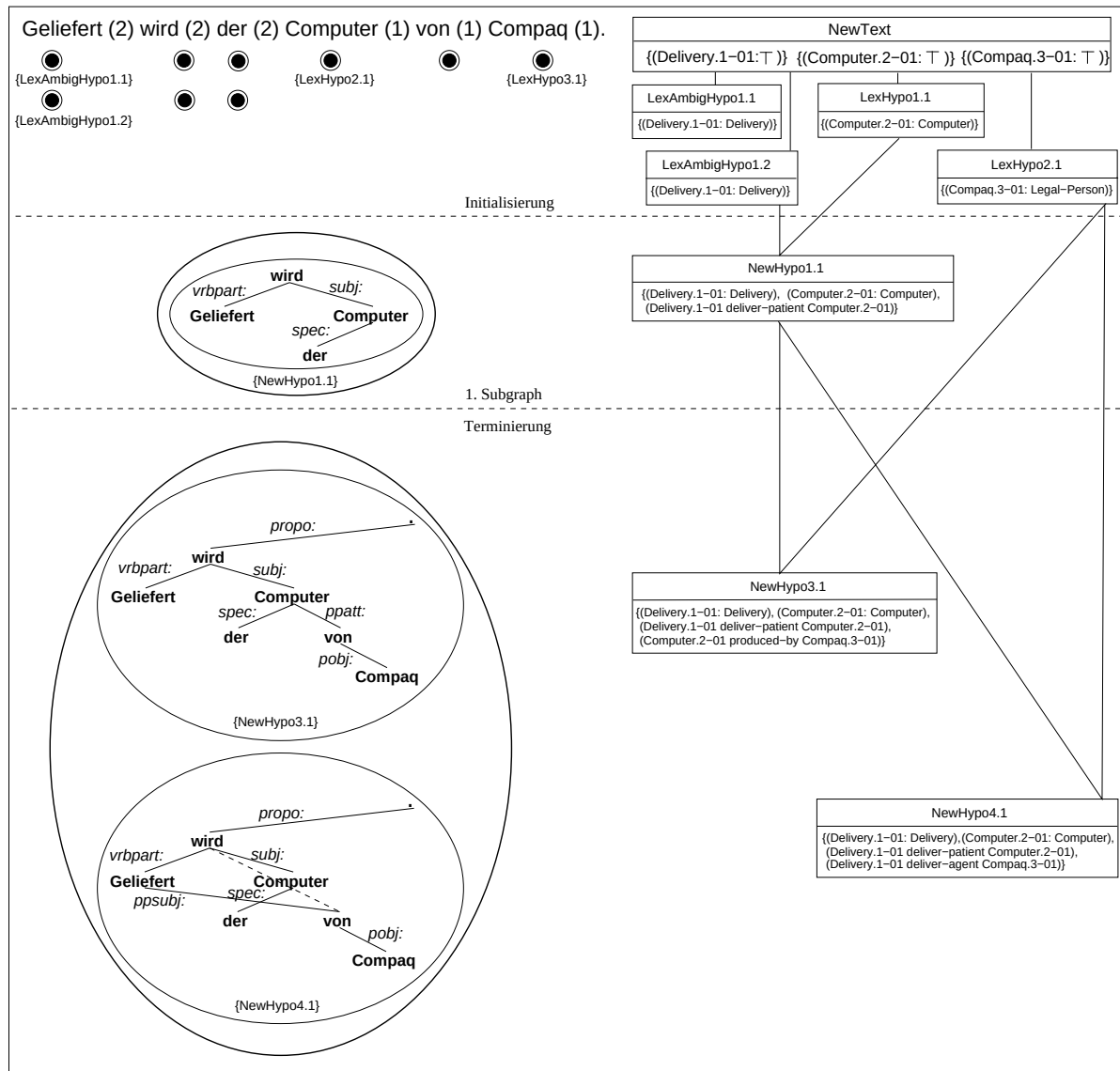


Abbildung 4.16: Repräsentation syntaktischer Ambiguitäten durch Kontexte

Interessant wird die Interaktion von Textwissensbasis und Syntax zu dem Zeitpunkt in der Analyse des Eingabesatzes, an dem die Präpositionalphrase “*von Compaq*” versucht, ihren syntaktischen Kopf zu bestimmen. Es ergeben sich zwei Anbindungsmöglichkeiten im Vorsatz, eine an das Verb “*geliefert*” und eine an das Nomen “*Computer*”. Eine Instantiierung des jeweiligen semantischen Interpretationsschemas für Präpositionalphrasen ist Bestandteil der Prüfung der Wohlgeformtheitsbedingungen, die in den Valenzen *ppsubject* bzw. *ppattribute* gebündelt sind. Weil Compaq eine Instanz des Konzepttyps LEGAL-PERSON ist (wie aus dem Initialkontext LEXHYPO2.1 ersichtlich), gibt es für beide angefragten syntaktischen Relationierungen eine gültige, aber auch inhaltlich divergierende Interpretation. Aus den statischen Constraints für die Präpositionalsubjekt-Lesart *ppsubject* ergibt sich deshalb, daß die Relationierung mit dem Verb ausschließlich in Form eines AGENT der ACTION einer DELIVERY zu interpretieren ist. Im Fall der syntaktischen Bindung der Präpositionalphrase an “*Computer*” berechnet die semantische Interpretation die Relation PRODUCED-BY. Die jeweils berechnete Relationierung wird in einem eigenen Kontext vorgenommen, der als Kindkontext unter den mit den Phrasenaktoren verbundenen Kontexten angelegt wird. Es kommen also folgende Assertionen hinzu (vgl. Abb. 4.16, unten):

(COMPUTER.2-01 PRODUCED-BY COMPAQ.3-01)_{NewHypo3.1}

(DELIVERY.1-01 DELIVER-AGENT COMPAQ.3-01)_{NewHypo4.1}

Diese Hypothesen über die Bedeutung des Eingabesatzes sind wiederum den neu erzeugten Phrasenaktoren mit den ambigen Strukturbeschreibungen für das PP-Attachment dadurch bekannt, daß jedem dieser Phrasenaktoren die Menge seiner ihn inhaltlich repräsentierenden Kontexte beigeordnet ist (vgl. Abb. 4.16, links unten). Das Ergebnis der Interpretation eines Einzelsatzes findet sich dann in den terminalen Kontexten, die disjunkte Hypothesen über den Satzinhalt darstellen.

Da mitunter – insbesondere im Fall einer unvollständigen syntaktisch-konzeptuellen Überdeckung des Eingabesatzes – eine größere Anzahl an terminalen Kontexten erzeugt wird, stellt sich innerhalb der Textanalyse das Selektionsproblem. Zu klären ist also die Frage, welche der berechneten interpretatorischen Alternativen im nachfolgenden Satz weiterverfolgt wird. Da der Nachfolgesatz aufgrund der unterschiedlichen Ebenen der Resolution anaphorischer Phänomene mit dem “besten” Ergebnis aus dem Vortext ausgestattet werden sollte, stellt die Kontextselektion am Satzende einen wichtigen Faktor bei der Textanalyse dar. Im Unterkapitel 4.3.8 wird dieses Problem mit einem pragmatischen Lösungs-

vorschlag aufgegriffen.

4.3.8 Kontextselektion am Satzende und Interaktion mit der Resolution satzübergreifender Phänomene

Dadurch, daß SYNDIKATE referentielle Bezüge oberhalb der Satzebene auflöst, stellen sich zwei weitere Fragen in Zusammenhang mit der Ambiguitätsproblematik. Erstens müssen am Satzende diejenigen Interpretationen weiterverfolgt werden, die mit großer Wahrscheinlichkeit valide Interpretationen des Eingabesatzes beinhalten (Selektionsproblem). Allein schon aufgrund der Kombinatorik der verschiedenen Ambiguitätsebenen können aus Komplexitätsgründen nicht alle Lesarten weiterverfolgt werden.²¹ Zweitens entstehen durch die Resolution anaphorischer Bezüge interpretatorische Abhängigkeiten zum Vortext, die in eine weitere Klasse von Ambiguitäten münden, nämlich *textuelle* Ambiguitäten. Beim momentanen Stand der Implementation tritt diese Ambiguitätsklasse jedoch nicht auf, da resolvierte anaphorische Bezüge die Löschung des Antezedenten aus der Centering-Liste auslösen. Dadurch kann jeder Antezedent nur einmalig abgerufen werden, ein Backtracking über die Centering-Liste ist nicht möglich. Im gegebenen Anwendungskontext fokussiert die Problemstellung daher allein auf die Hypothesenselektion an jedem Satzende einer Folge von Eingabesätzen aus einem Text.

Die gültigen Interpretationen eines Eingabesatzes finden sich grundsätzlich in den terminalen Kontexten der die Textwissensbasis repräsentierenden Kontexthierarchie. Für die weitere Analyse des Eingabetexts sind die terminalen Kontexte zu selektieren, die möglichst vollständige (Teil-)Berechnungsergebnisse des Parsers inhaltlich reflektieren. Damit scheiden jene terminalen Kontexte aus, die syntaktisch nicht validierte Analyseergebnisse darstellen, wie etwa mit nicht weiterverfolgten Wortklassenhypothesen verbundene Initialkontexte, und solche, die aufgrund des präferentiellen und partiellen Parsingansatzes nicht weiterberechnet wurden.

Der Selektionsmechanismus für den besten Interpretationskontext setzt auf zwei Ebenen an und vollzieht sich ausschließlich auf der Seite der Textwissensbasis, ohne den Parser zu involvieren. Zunächst wird die Überdeckung innerhalb des Dependenzgraphen berücksichtigt, welche sich unmittelbar in der Pfadlänge vom Wurzelkontext NEWTEXT zu einem

²¹Legt man eine mittlere Textlänge von ca. 20 Sätzen zugrunde und geht in einem idealisierten Szenario von nur 2 verbleibenden Lesarten pro Satz aus, dann verblieben am Textende 2^{20} inhaltliche Alternativen.

terminalen Kontext widerspiegelt, denn darin drückt sich die Anzahl der in einem Kontext eingeschlossenen minimalen interpretierbaren Subgraphen aus. Darüber hinaus findet jedoch auch die *Qualität* der semantisch-konzeptuellen Interpretation Eingang in die Bewertung.

Die formale Grundlage für die Bewertung eines Kontexts besteht darin, Kontexte als formale Objekte aufzufassen, über die Meta-Aussagen getroffen werden können (McCarthy & Buvač, 1998). Damit kann die Bewertung in Form eines numerischen Gewichts der Beschreibung eines Kontexts über eine Relation hinzugefügt werden. Gewichtet wird dies durch eine Bewertungsfunktion $w : h \rightarrow \mathbb{Q}_0^+$ mit $h \in \mathcal{H}$, welche die Qualität einer Interpretation für einen beliebigen Kontext in einem positiven numerischen Wert ausdrückt. Die Bewertungsheuristiken, welche bei der Gewichtung herangezogen werden, befinden sich noch im Stadium der Entwicklung. Jedoch lassen sich mit der aktuellen Vorgehensweise gute Näherungen erzielen. Darüber hinaus sei angemerkt, daß sich mit diesem flexiblen Mechanismus selbst aus kognitionspsychologischer Sicht sinnvolle Bindungspräferenzen, wie etwa minimal attachment oder right closure, durch unterschiedliche Gewichtung von Präpositionalphrasen modellieren ließen.

Die aktuell zum Einsatz kommende Gewichtungsfunktion basiert auf einem recht einfachen Bewertungsschema. Das Gewicht eines beliebigen Kontexts h im Kontextgraphen eines Eingabetexts ermittelt sich aus dem Maximum des Gewichts aller *direkten* Elternkontexte, also aller Kontexte, die mit h über *direct-supercontext-of* relationiert sind, zuzüglich des Gewichts für die semantische bzw. konzeptuelle Interpretation in h . Für die Bewertung eines einzelnen Kontexts ist festgelegt, daß das Gewicht der initialen Textwissensbasis NEW-TEXT und von Initialkontexten gleich 0 ist. Die semantische Interpretation einer “direct linkage” und einer konzeptuellen Interpretation werden jeweils mit 1 bewertet. Schließlich erhält eine erfolgreich resolvierte satzübergreifende Anapher ebenso eine 2 als Wert wie eine semantische Interpretation einer “mediated linkage”.²²

Jedoch ist die prinzipielle Verfügbarkeit einer Methode, Interpretationen bewerten zu können, zu trennen von der nur empirisch bestimmbaren Frage, wie die einzelnen Interpretationsschritte zu gewichten sind. Zu letzterem fehlt aufgrund der Tatsache, daß hier wissenschaftliches Neuland betreten wird, eine ausreichend fundierte Datengrundlage, so

²²Da nicht auszuschließen ist, daß aufgrund von Erfahrungen mit dem Textverstehenssystem auch nicht natürliche Zahlen in die Gewichtung eingehen, liegt der Wertebereich der Gewichtungsfunktion w allgemeiner auf den positiven rationalen Zahlen.

daß die festgelegten Werte zwar der Erfahrung mit den Analyseergebnissen von SYNDIKATE entspringen, jedoch nur grobe Näherungen bedeuten und weiter zu verfeinern sowie experimentell zu evaluieren sind.

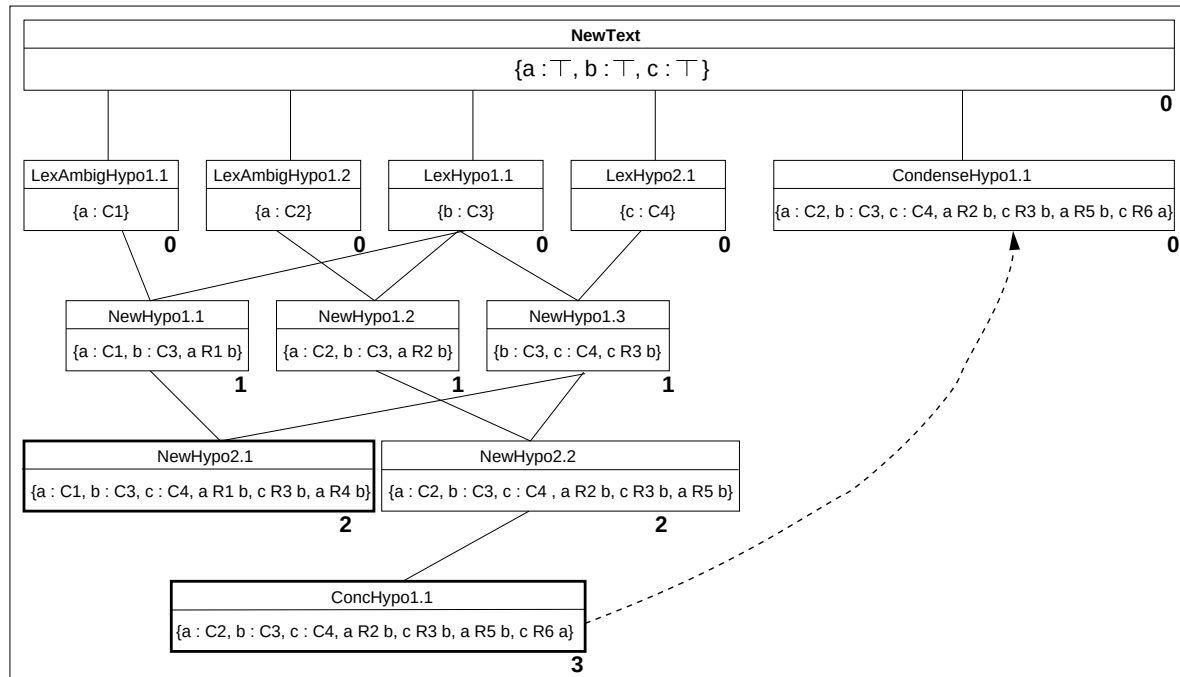


Abbildung 4.17: Kontextselektion und Kontextkondensierung

Zugleich mit der Selektion der besten Lesarten für einen Eingabesatz wird die Anzahl der Kontexte dadurch reduziert, daß die beibehaltenen Interpretationskontexte kondensiert und alle anderen nicht mehr benötigten Kontexte gelöscht werden. Die Kondensation eines Kontexts besteht aus der Generierung eines neuen Initialkontexts, in den alle Assertionen des zu kondensierenden Kontexts eingestellt werden.

Zur Illustration des Verfahrens enthält Abb. 4.17 ein schematisiertes Beispiel. In der Initialisierungsphase werden die Instanzen a , b und c getypt (Konzepte $C1$, $C2$, $C3$ und $C4$) in die Kontexte LEXAMBIGHYPO1.1, LEXAMBIGHYPO1.2, LEXHYPO1.1 sowie LEXHYPO1.2 eingestellt. Im Verlauf der Satzanalyse generiert SYNDIKATE eine Reihe neuer Kontexte, um die Ergebnisse der semantischen Interpretation zu repräsentieren. Diese sind als Relationierungen zwischen Instanzen in Form von assertionalen Axiomen jeweils den Kontexten hinzugefügt. Im Kontext NEW-HYPO2.2 führt das Hinzufügen eines assertionalen Axioms darüber hinaus zu einer konzeptuellen Interpretation. Der Kontext

CONCHYPO1.1 beinhaltet daher als Hypothese über den Satzinhalt ein weiteres assertionales Axiom. Am Satzende gibt es nur zwei terminale Kontexte, nämlich die graphisch durch fett gerahmte Rechtecke hervorgehobenen Kontexte NEW-HYPO2.1 und CONCHYPO1.1. Die Selektion der besten Lesart geschieht nun auf der Grundlage der Gewichte an den Kontexten, die in Abb. 4.17 als Ziffern an den Kontexten angefügt sind. Da bei den beiden terminalen Kontexten der Kontext CONCHYPO1.1 das größere Gewicht aufweist, wird er als Interpretationsergebnis selektiert und anschließend kondensiert. Der kondensierte Kontext CONDENSEHYPO1.1 besteht aus einer exakten Kopie der Assertionen in CONCHYPO1.1, ist aber im Gegensatz zu diesem ein direkter Subkontext von NEWTEXT. Alle Kontexte des Eingabesatzes können nun gelöscht werden, einzig CONDENSEHYPO1.1 bleibt als Surrogat des Eingabesatzes erhalten.

Die Resolution anaphorischer Phänomene auf der Textebene setzt nun genau an diesen kondensierten Kontexten an. Alle Diskursobjekte, die in einem Satz referentiell aufgenommen werden können, befinden sich in dem zum vorangegangenen Satz gehörigen kondensierten Kontext.

Tritt nun innerhalb eines Satzes ein textuelles Phänomen auf, so muß geprüft werden, ob der mögliche anaphorische Bezug zum Vortext in eine gültige Interpretation aufzulösen ist.²³ Im Gegensatz zum bisher dargestellten, auf Satzebene operierenden Kontextmechanismus erfolgt die kontextuelle Anknüpfung durch eine Verbindung mit dem Vortext außerhalb des aktuellen Satzes. Bedingt durch die nur zwischen unmittelbar aneinandergrenzenden Sätzen arbeitende Anaphernresolution finden sich die potentiellen Antezedenten im kondensierten Kontext des vorangegangenen Satzes. Der Zugriff auf die Antezedenten erfolgt über die in den Centeringlisten abgelegten Tupel bestehend aus dem Instanzidentifizier in der Textwissensbasis und der Wortoberfläche im Vorsatz (welche im Fall einer textuellen Ellipse allerdings nicht realisiert ist). Kann die Anapher nun erfolgreich aufgelöst werden, dann muß dies in der Textwissensbasis eine Transaktion in Form einer Assertion nach sich ziehen. Um aber die Diskursobjekte aus dem Vortext verfügbar zu machen, wird ein neuer Kontext generiert, der einerseits den zum Vortext gehörenden kondensierten Kontext und andererseits den satzinternen Kontext mit dem anaphorischen Ausdruck als direkten Superkontext hat.

²³Auf das Anfrageverhalten der beteiligten Akteure soll an dieser Stelle nicht weiter eingegangen werden (vgl. hierzu Strube (1997)), da im Zusammenhang dieser Arbeit allein das Kontextmanagement relevant ist.

Die fortlaufend durchgeführte Resolution extrasententieller textueller Phänomene –wie in Kap. 3.2.5 beschrieben– durch referentielle Verknüpfung von Einzelsatzanalyse resultiert am Textende in einer Menge an terminalen Kontexten, die zugleich auch kondensierte Kontexte sind. Zur abschließenden inhaltlichen Repräsentation eines Eingabetexts werden die in den kondensierten Kontexten eingeschlossenen Mengen an assertionalen Axiomen vereinigt. Interpretatorische Alternativen sind auch auf dieser Ebene als disjunkte Mengen assertionaler Axiome repräsentiert. Allerdings impliziert das Ambiguitätsproblem das potentielle Vorhandensein mehrerer Textwissensbasen für einen einzigen Eingabetext. Da jedoch jede dieser Lesarten durch die semantische Interpretation syntaktischer Strukturen ebenso wie durch die textuelle Progression sanktioniert ist, müssen diese Mehrdeutigkeiten bezüglich des Textinhalts in Kauf genommen werden.

Abschließend beurteilt weist der verwendete methodische Ansatz zur Repräsentation von Ambiguitätsphänomenen, die im Laufe einer Textanalyse entstehen, eine Reihe von Vorteilen auf. Zunächst bestimmt der Ursprung einer Ambiguität den Repräsentationsort, wodurch die angestrebte Modularisierung, d.h. die Trennung der Repräsentationsschichten und Wissensquellen von der Wortebene bis hin zur Textebene realisiert ist. Insbesondere interferieren semantisch-lexikalisch bedingte Ambiguitäten nicht mit der syntaktischen Analyse. Durch die Verwendung eines in die terminologische Logik eingebetteten Kontextmechanismus erhält das Verfahren die notwendige formale Fundierung. Einerseits dienen Kontexte als Hypothesenräume, in denen unterschiedliche inhaltliche Repräsentationen konsistent abgebildet sind, andererseits werden bei der Selektion der besten Lesart eines Satzes Kontexte als formale Objekte beim Metareasoning verwendet. Auf der Grundlage der in den Kontexten eingeschlossenen assertionalen Axiomen transportieren diese Kontexte den textuellen Kontext, der zur Disambiguierung Verwendung findet, weil nur solche Lesarten weiterverfolgt werden, die konsistent um zusätzliche Interpretationen zu erweitern sind.

Kapitel 5

Wissensrepräsentationsformalismen und Domänenwissen

Die Inferenzprozesse, auf denen SYNDIKATE während des Textverstehens aufsetzt, sind stark wissensbasiert und verwenden gleichermaßen Allgemeinwissen und domänenspezifisches Wissen. Das semantische Filtern etwa basiert auf Pfadmustern entlang von Typbeschränkungen, die für Relationen an Konzeptbeschreibungen definiert sind. Die Resolution textueller Phänomene benutzt bei nominalen Anaphern unter anderem taxonomisches Wissen. Die semantische Interpretation für Präpositionen gründet sich auf Abstraktionen in der Relationshierarchie, um die Suche nach gültigen Relationspfaden zu restringieren. Diese repräsentativen Beispiele mögen den Stellenwert sauber modellierter und logisch konsistenter Wissensbasen illustrieren. Jedoch beschränkt sich die Funktion der Wissensbasen nicht auf die reine Repräsentation a priori bereitgestellten Wissens. Während der Textanalyse erweitern die extrahierten assertionalen Axiome fortlaufend die anfänglich aus einer Kopie der IT-KB oder MED-KB bestehende initiale Textwissensbasis.

Im Bereich Wissensrepräsentation gibt es zwei thematische Schwerpunkte, die beim Textverstehen von Bedeutung sind. Erstens: die Wissensrepräsentationsformalismen, die sowohl determinieren, welches Wissen überhaupt modellierbar ist, als auch eine obere Grenze dafür festlegen, welche Tiefe des Textverstehens erreichbar ist. Und zweitens: das Domänenwissen hinsichtlich seiner konzeptuellen Abdeckung und seiner Granularität.

Formal-fundierte und logik-basierte Wissensrepräsentationssprachen bieten ein Inventar an Ausdrucksmitteln an, um Beschreibungen von Weltwissen adäquat repräsentieren zu können. Auf der Ebene der sie implementierenden Wissensrepräsentationssysteme bringen

die Komplexitätseigenschaften von Algorithmen (etwa zur Konsistenzprüfung innerhalb einer Wissensbasis) jedoch Einschränkungen mit sich, die sich unmittelbar auf das ontologische Design auswirken. Unter der doppelten Perspektive, welche Problemfelder es bei der Ausdrucksmächtigkeit von Wissensrepräsentationssprachen gibt und wie die Expressivität die Verstehensleistung affiziert, führt das Kapitel 5.1 eine Reihe von ontologischen Beschreibungsebenen ein. Ergänzend dazu befaßt sich Kapitel 5.2.1 mit den konkreten Anforderungen an die Wissensmodellierung, die sich aus dem Anwendungskontext in SYNDIKATE ergeben, und legitimiert den in SYNDIKATE implementierten ontologischen Entwurf.

Da der enorme quantitative und qualitative Bedarf an ontologischem Wissen, der sich aus dem Ansatz in SYNDIKATE zum tiefen und breiten Textverstehen ergibt, nicht ausschließlich durch eigene Modellierungsleistungen abzudecken ist, gibt es gute Gründe, die Arbeiten von Gruppen mit langjähriger Erfahrung im Ontological Engineering zu analysieren. Zum einen können die überaus zeitaufwendigen Reflexionen über sehr abstrakte Konzepte und deren taxonomische Beziehungen zueinander durch Anleihen verkürzt werden. Zum anderen lassen sich bereits existierende Wissensrepositorien unterschiedlichster Provenienz mit Hilfe (semi-)automatischer Verfahren in die eigenen Wissensbasen übertragen (Kapitel 5.3).

5.1 Wissensrepräsentationssprachen und Wissensrepräsentationssysteme

Die Ausdrucksmächtigkeit eines formalen Systems, dessen Komplexitätseigenschaften sowie die sich aus einer Applikation –wie etwa SYNDIKATE– heraus ergebenden Anforderungen an die Wissensrepräsentation formen ein magisches Dreieck, dessen Eckpunkte sich gegenseitig beeinflussen. Dabei gilt es, im Rahmen einer Anwendung insbesondere die Ausdrucksmächtigkeit eines formalen Systems mit seinen Komplexitätseigenschaften auszutarieren. Dieses Kapitel möchte die ersten beiden Ecken dieses Dreiecks diskutieren, während im nachfolgenden Kap. 5.2 ein Minimalkatalog für die Expressivität eines Wissensrepräsentationssystems innerhalb unseres Textverstehenssystems aufgelistet ist. Obwohl eine größere Anzahl an Wissensrepräsentationsformalismen existiert (Reimer, 1991), konzentriere ich mich bei der Darstellung auf die Beschreibungslogik, weil diese einerseits

dem in SYNDIKATE verwendeten Wissensrepräsentationssystem LOOM zugrunde liegt, andererseits aber auch bei der Modellierung großer Ontologien eine zunehmend wichtigere Rolle spielt (vgl. Kap. 5.3.2).

Mit der Festlegung auf ein formales System zur Wissensrepräsentation schreibt man zugleich auch fest, welches Wissen überhaupt darstellbar ist. Auch die terminologische Logik weist Beschränkungen bezüglich des darstellbaren Wissens auf. Beispielsweise stellten Haimowitz et al. (1988) bei der Portierung eines Systems zum diagnostischen Reasoning in das terminologische Wissensrepräsentationssystem NIKL (Schmolze & Mark, 1991) fest, daß die Expressivität dieser Konzeptsprache für das Knowledge Engineering –zumindest im betrachteten Anwendungskontext– unzureichend ist. Ausschlaggebend für die Ausdrucksmächtigkeit einer terminologischen Wissensrepräsentationssprache sind neben der Menge an konzept- und rollenformenden Operatoren (und syntaktischer Beschränkungen bei deren Verwendung) temporale Konzept- bzw. Relationsbeschreibungen sowie die Menge an verfügbaren Quantoren.

Im folgenden sollen diese Ebenen im Kontext des Knowledge Engineering wie auch der automatischen Analyse natürlicher Sprache thematisiert werden. Anschließend folgen einige Bemerkungen zu existierenden Wissensrepräsentationssystemen, da eine Anwendung wie SYNDIKATE auf einer tatsächlich implementierten terminologischen Wissensrepräsentationssprache aufsetzen muß.

Terminologische Sprachen variieren stark in ihrer Ausdrucksmächtigkeit, je nachdem welches Inventar an konzept- und rollenformenden Operatoren sie zur Verfügung stellen (eine Übersicht hierzu findet sich in Woods & Schmolze (1992)).¹ Levesque & Brachman (1987) analysieren den Zusammenhang zwischen Ausdrucksmächtigkeit, gegeben durch die Menge an konzept- und rollenformenden Operatoren, und der Traktabilität der hierdurch definierten Sprachen, mit denen Wissen repräsentiert werden kann. Im Bereich der terminologischen Logik bestehen die von der deduktiven Komponente zu leistenden Schlußfolgerungen darin, in der T-Box Subsumptionsbeziehungen zwischen Konzepten sowie die Erfüllbarkeit von Konzeptdefinitionen abzuleiten (Donini et al., 1996). In seiner formalen Untersuchung zur Berechnungskomplexität der Konzeptklassifikation kommt Nebel (1990b) zunächst zu dem Ergebnis, daß terminologische Logik im schlechtesten Fall – der indessen äußerst selten eintritt – intractabel ist. Jedoch läßt sich die Worst-Case-Komplexität durch

¹Weitere Übersichtsarbeiten zu unterschiedlichen terminologischen Sprachen stammen außerdem von Schmidt-Schauß & Smolka (1991) und Donini et al. (1997).

zwei verschiedene Strategien reduzieren. Zum einen kann man die Ausdrucksmächtigkeit der terminologischen Sprache beschränken (Donini et al., 1991), zum anderen können unvollständige Klassifikationsalgorithmen zum Einsatz kommen.

Wenn es darum geht, die Ausdrucksmächtigkeit terminologischer Logiken zu erweitern, dann kann auf unterschiedlichen Ebenen angesetzt werden. Zunächst betreffen Erweiterungen die Syntax der Beschreibungssprache selbst oder die Menge an konzept- und rollenformenden Operatoren sowie die Ausdrucksmittel, mit denen Konzepte mit Hilfe von Relationen definiert sind. Zur Syntax gehört zum Beispiel die Erlaubnis zyklischer Konzeptdefinitionen, im Bereich der bereitgestellten Operatoren die Disjunktion, die Konjunktion oder etwa die Negation. Die Sprache $\mathcal{ALC}$ (Schmidt-Schauß & Smolka, 1991), die eine Teilmenge der in Kapitel 3.2.2 vorgestellten Sprache $\mathcal{CDL}$ ist, kann durch folgende abstrakte Syntax definiert werden (sie dient in den folgenden Ausführungen als Referenzsprache):

$$C, D \rightarrow A \mid C \sqcup D \mid C \sqcap D \mid \neg D \mid \exists R.C \mid \forall R.C.$$

Durch Integration weiterer Sprachelemente – wie etwa numerischer Beschränkungen an Rollen – entstehen ausdrucksmächtigere Sprachen zu $\mathcal{ALC}$.

Eine völlig andere Ebene der Erweiterung terminologischer Logiken betrifft das Folgern über spatiale bzw. temporale Beziehungen und das über die Existenz- und Allquantifikation hinausreichende Folgern, wie es die Analyse natürlicher Sprache erfordert. Die hierin eingeschlossenen Schlußfolgerungen sind zunächst nicht auf die den terminologischen Logiken eigenen deduktiven Reasoningmechanismen, nämlich Konzeptsubsumption und Erfüllbarkeit von Konzeptdefinitionen in einer T-Box, zurückführbar.

Grundsätzlich bieten sich zwei Lösungsansätze, um sprachliche Erweiterungen in terminologische Logik zu integrieren. Erstens kann mit den Mitteln einer gegebenen terminologischen Sprache – etwa durch Reifikation oder “Kunst”-Konzepte – eine Metaebene eingeführt werden, die die erforderlichen Schlußfolgerungen auf Subsumption und Erfüllbarkeit zurückführt (in diesen Bereich fallen etwa die Spracherweiterungen des Lernalgorithmus in SYNDIKATE (Schnattinger & Hahn, 1998) durch Reifikation (MacGregor, 1993) oder die in Kapitel 5.2.1.3 beschriebenen Erweiterungen zu Partonomien durch sogenannte Triplets (Hahn et al., 1999c)). Zweitens kann eine terminologische Sprache durch die Einführung neuer Operatoren erweitert werden. Jedoch wirkt sich gerade letzteres auf die formalen Eigenschaften wie auf die Komplexitätseigenschaften der Sprache aus. Im Fall einer Metabeschreibung spreche ich von einer *schwachen* Semantik, mit der ursprünglich nicht in der terminologischen Sprache beinhaltete logische Bezüge durch Metakonstrukte repräsen-

tiert werden. Im Fall einer tatsächlichen Erweiterung der Sprache um zusätzliche formale Konstrukte mit einer klar definierten formalen Semantik spreche ich von einer *starken* Semantik. Im nachfolgenden diskutierte ich kurz verschiedene Aspekte terminologischer Sprachen und ordne diese den angeführten Ebenen zu.

5.1.1 Terminologische Zyklen

Nebel (1991) unterscheidet drei Arten von terminologischen Zyklen. Diese Unterscheidung wird ausschließlich auf der Grundlage der syntaktischen Form von Konzeptdefinitionen getroffen, also unabhängig davon, ob bestimmte zirkuläre Definitionen ontologisch überhaupt Sinn machen. Unter einem terminologischen Zyklus versteht man zunächst das Auftreten eines Konzeptsymbols C_1 auf der linken wie auf der rechten Seite eines terminologischen Axioms, wobei das Symbol C_1 auf der rechten Seite im Skopus eines Allquantors oder eines Existenzquantors steht. Daß diese Form der zirkulären Konzeptdefinition sinnhaft in rekursiven Konzeptbeschreibungen zu verwenden ist, zeigt die nachfolgende Definition:

$$\text{PHYSICAL-OBJECT} \doteq \text{OBJECT} \sqcap \exists \text{HAS-PHYSICAL-PART. PHYSICAL-OBJECT}$$

Ein physikalisches Objekt definiert sich dadurch, daß es als physikalische Bestandteile wiederum physikalische Objekte besitzt.

Neben dieser Form der *restriction-circular concepts* gibt es sogenannte *component-circular concepts* (Nebel, 1991), bei denen zwei Konzepte C_1 und C_2 sich wechselseitig in der Konzeptdefinition referenzieren, ohne dabei im Skopus eines Quantors zu stehen. Im folgenden Beispiel zu dieser Form zyklischer Axiome definieren sich die Konzepte MAN und MALE-HUMAN, indem sie einander referenzieren:

$$\begin{aligned} \text{MAN} &\doteq \text{HUMAN} \sqcap \text{MALE-HUMAN} \\ \text{MALE-HUMAN} &\doteq \text{HUMAN} \sqcap \text{MAN} \end{aligned}$$

Ontologisch betrachtet scheint diese – in terminologischen Sprachen, die Zyklen zulassen, syntaktisch korrekte – Form der *component-circular* Konzeptdefinition indessen keinen Sinn zu machen (Nebel, 1991).

Als dritte Form zyklischer Konzeptdefinitionen gibt es die *circular object structures*.

Eine zirkuläre Objektstruktur liegt dann vor, wenn zwei Konzepte, C_1 und C_2 sich in ihren Definitionen wechselseitig referenzieren und dabei das jeweils andere Konzept im Skopus eines Allquantors oder eines Existenzquantors steht. Die Relationierung erfolgt dabei in C_1 über eine Relation R_1 und in C_2 über deren Inverse R_1^- .

STORAGE $\doteq$ HARDWARE $\sqcap \exists$ HAS-CAPACITY. STORAGE-CAPACITY
STORAGE-CAPACITY $\doteq$ CAPACITY $\sqcap \exists$ CAPACITY-OF. STORAGE

Die Hardware eines physikalischen Speichers hat als Eigenschaft eine Speicherkapazität, diese wiederum definiert sich als die Kapazität eines Speichers.

In terminologischen Logiken wirken sich Zyklen auf zwei Ebenen aus. Zunächst beeinflussen terminologische Zyklen die Komplexität der Klassifikationsalgorithmen dahingehend, daß in ausdrucksmächtigeren Sprachen die Subsumption von Konzepten unentscheidbar wird (Nebel, 1991). Darüber hinaus ergibt sich das Problem, welche Semantik zyklischen Definitionen bei der Klassifikation zu geben ist, da die modelltheoretische Semantik in einigen Fällen keine eindeutige Interpretation für zyklische Konzeptdefinitionen liefert (vgl. hierzu Baader (1990)).

Einen Zwischenweg, um Ausdrucksmächtigkeit und Komplexität auszutarieren, gehen Buchheit et al. (1998), die für eine Aufspaltung des terminologischen Systems in sogenannte “*Schemata*” und “*Views*” plädieren. In den *Schemata* sind ausgehend von den ontologischen Regularitäten abstraktere Top-Level-Konzepte beschrieben. Die Konzeptbeschreibungssprache für *Schemata* läßt zyklische Definitionen zu, Konzeptdefinitionen beinhalten indessen grundsätzlich nur notwendige Bedingungen (d.h., es handelt sich um primitive Konzepte). Demgegenüber weisen Konzeptdefinitionen in den *Views* notwendige und hinreichende Bedingungen auf. Dafür sieht die Konzeptbeschreibungssprache für *Views* keine zyklischen Definitionen vor. Außerdem stellt die Konzeptbeschreibungssprache für *Schemata*, $\mathcal{SL}$ (Buchheit et al., 1994), weniger konzeptformende Operatoren zur Verfügung als die für *Views* verwendete Sprache $\mathcal{ALCN}\mathcal{R}$ (Hollunder et al., 1990). Eine Wissensbasis ist folglich ein Tupel $\langle \mathcal{S}, \mathcal{W} \rangle$, wobei $\mathcal{S}$ ein $\mathcal{SL}$ -Schema und $\mathcal{W}$ eine in $\mathcal{ALCN}\mathcal{R}$ gegebene Weltbeschreibung darstellt.

In den SYNDIKATE-Wissensbasen wären terminologische Zyklen der ersten und der dritten Form ein wünschenswertes Sprachkonstrukt der zugrundeliegenden Wissensrepräsentationssprache bzw. des zugrundeliegenden Wissensrepräsentationssystems, weil zy-

klische Definitionen in der Tat eine wichtige Rolle bei der Wissensmodellierung spielen (Kaczmarek et al., 1986). Beispielsweise wären nahezu alle Konzepte, die Maßeinheiten repräsentieren (immerhin ca. 60), in einer strukturzyklischen Form zu definieren. Bedauerlicherweise unterstützt das verwendete Wissensrepräsentationssystem LOOM keine zyklischen Definitionen, wenn in den Konzept- und Relationsdefinitionen ausschließlich hinreichende *und* notwendige Bedingungen auftreten. Die dadurch reduzierten Möglichkeiten beim Matchen von Instanzen, d.h. der Inferenz, ob eine Instanz zu einem Konzept gehört oder nicht, hat unmittelbare Rückwirkungen auf die Textverstehensleistung des Gesamtsystems (vgl. insbesondere das nachfolgende Kapitel 5.2.1).

5.1.2 Transitivität von Relationen in terminologischer Logik

Neben der Erweiterung terminologischer Sprachen um Formen zyklischer Konzeptdefinitionen gehört die Transitivität von konzeptuellen Relationen zu den sprachinhärenten Erweiterungen, da neben der Konzeptklassifikation und den Berechnungen, ob Konzeptdefinitionen einer gegebenen Terminologie erfüllbar sind, kein erweitertes Reasoning eingeführt wird. Allerdings verlassen um transitive Rollen erweiterte Konzeptbeschreibungssprachen die Ebene der Logik 1. Ordnung (Baader, 1991). Im Prinzip lassen sich transitive Relationen auch mithilfe zyklischer Definitionen des *restriction-circular-concept*-Typs ausdrücken, wie anhand der im vorangegangenen Kapitel 5.1.1 enthaltenen Definition für PHYSICAL-OBJECT unmittelbar zu überprüfen ist. In diesem Sinn stellen Konzeptsprachen mit transitiven Relationen eine Alternative zu Konzeptsprachen mit terminologischen Zyklen dar (Baader, 1991).

Die Frage, inwieweit transitive Rollen zur Konzeptdefinition in die terminologische Konzeptbeschreibungssprache zu integrieren sind, wird wegen der Komplexitätseigenschaften (Sattler, 1996) kontrovers diskutiert und läßt sich letztlich nur auf der Grundlage konkreter Modellierungsanforderungen einer Domäne beantworten. Zur Illustration des Problems mag folgender (natürlich stark vereinfachter Sachverhalt) dienen: ein Computer sei definiert als eine Hardware, von der mindestens ein Teil ein Prozessor ist. Formal in Beschreibungslogik notiert als:

$$\text{COMPUTER-SYSTEM} \doteq \text{HARDWARE} \sqcap \exists \text{ HAS-PART. CPU}$$

Jedoch muß das Konzept CPU im ontologischen Entwurf nicht direkt mit COMPUTER-SYSTEM relationiert sein. In einer feinkörnigeren Modellierung könnte eine CPU indessen Teil eines MOTHERBOARD sein, das wiederum Teil eines Computers ist usw. In letzter Konsequenz weiß der Knowledge Engineer zum Zeitpunkt der Modellierung eines Konzepts nicht, wieviele PART-OF Relationen letztendlich zu traversieren sind, da der notwendige Granularitätsgrad keine a priori determinierbare Größe darstellt. Folglich wäre es eleganter das Konzept COMPUTER-SYSTEM in Anlehnung an Sattler (1996) wie folgt als ein Konzept zu definieren, das entweder direkt das physikalische Teil CPU hat oder über eine Anzahl von HAS-PART Relationen mit CPU verbunden ist:

$$\text{COMPUTER-SYSTEM} \doteq \text{HARDWARE} \sqcap (\exists \text{ HAS-PART: CPU} \sqcup \exists \text{ HAS-PART*}: \text{CPU})$$

Analoge Beispiele existieren in größerer Anzahl in der Medizin. Ein Hirntumor ließe sich zum Beispiel dadurch definieren, daß irgendein anatomisches Teil des Gehirns oder aber eine Gewebestruktur von krebsartigem Wachstum befallen ist. Die Verfügbarkeit transitiver Relationen ermöglicht auch in diesem Fall, das Konzept flexibel zu definieren, indem die befallene anatomische Struktur nicht explizit in der Konzeptdefinition des Ganzen referenziert werden muß (vgl. Horrocks & Sattler (1998b) und Horrocks & Sattler (1998a)).

Aus den angeführten Beispielen läßt sich folgern, daß eine Konzeptsprache mit transitiven Rollen grundsätzlich wünschenswert wäre. Der Trade-Off zwischen Berechnungskomplexität und Ausdrucksmächtigkeit wird deutlich, wenn man die konzept- und relationsformenden Operatoren der terminologischen Wissensrepräsentationssprache GAIL analysiert (Rector et al., 1997). Wegen der besonderen Relevanz der Transitivität von Relationen und des daran geknüpften Folgerns in der Medizin – wie sich deutlich in partonymischen Beziehungen in der Anatomie manifestiert (vgl. Kapitel 5.2.1.3) – implementiert GAIL als einziges terminologisches System transitive Relationen mit einem vollständigen Klassifikationsalgorithmus. Ein spezieller, in die Wissensrepräsentationssprache integrierter Operator (*specializes*) definiert die Transitivitätseigenschaft von Relationen. Jedoch gleicht GAIL diese punktuelle Spracherweiterung durch weitreichende Beschränkungen in der Ausdrucksmächtigkeit aus. Es gibt keine Disjunktion, keine Negation, keine Allquantifikation bei Relationsrestriktionen sowie keine numerischen Beschränkungen für die Anzahl von Füllern einer konzeptuellen Relation (Horrocks et al., 1996). Damit läßt sich deutlich illustrieren, daß aufgrund von Beschränkungen, die durch die Komplexitätseigenschaften

terminologischer Systeme auferlegt werden, seitens des Knowledge Engineering eine Auswahl zu treffen ist, welche Sprachkonstrukte in der eigenen Wissensbasis unerlässlich sind (vgl. hierzu Doyle & Patil (1991)). Letztendlich läßt sich diese Frage jedoch nur empirisch beantworten. Der in Kapitel 5.2.1 enthaltene Katalog beschreibt ein Minimum an benötigter Ausdrucksmächtigkeit für das Textverstehen auf der Grundlage der mit dem Textverstehenssystem SYNDIKATE gemachten Erfahrungen.

5.1.3 Temporal Reasoning in Beschreibungslogik

Da temporale Bezüge in Texten zweifelsohne eine Rolle spielen, muß die inhaltliche Repräsentation, die im Laufe der Textanalyse erzeugt wird, in der Lage sein, temporale Beziehungen abzubilden. In der Medizin sind zeitliche Relationen darüber hinaus auch zwischen Texten möglich, weil sich in der elektronischen Krankenakte der zeitliche Verlauf einer Patientengeschichte durch eine zeitlich geordnete Menge von Texten widerspiegelt.²

Prinzipiell gibt es die bereits angeführten beiden Möglichkeiten, um temporale Relationen zwischen Konzepten zu repräsentieren. Einerseits sind zeitliche Beziehungen mit den Mitteln einer terminologischen Sprache im Sinne einer schwachen Semantik beschreibbar. Andererseits können terminologische Sprachen im Sinne einer starken Semantik um modelltheoretisch fundierte temporale Schlußfolgerungsmechanismen erweitert werden.

Schmiedel (1990) und Schild (1993) führen in terminologische Sprachen temporale Quantoren ein, deren Semantik mit Hilfe eines diskreten Zeitmodells fixiert wird. Im Kontext temporaler Bezüge erhält der Subsumptionsbegriff eine umfassendere Bedeutung dadurch, daß zeitliche Inklusion bei der Konzeptklassifikation mitberücksichtigt wird. D.h., ein Zustand oder ein Ereignis mit einer kontinuierlichen zeitlichen Ausdehnung – wie etwa das Besitzen eines Objekts – subsumiert ein Konzept des gleichen Zustands bzw. Events, wenn die zeitliche Ausdehnung des letzteren vollständig in der des ersteren eingeschlossen ist. Aufbauend auf den temporalen Quantoren fügt Schmiedel (1990) seiner terminologischen Logik ein Netzwerk mit temporalen Relationen zur Konzeptbeschreibung hinzu, er kann jedoch keine Algorithmen zur Klassifikation für seine temporale terminologische

²Neben temporalen Bezügen spielen in Texten auch spatiale Relationierungen und Modalität eine Rolle, die in der logischen Repräsentation ihren Niederschlag finden sollte. In diesem Unterkapitel ist exemplarisch das temporale Reasoning herausgegriffen, um kritisch zu diskutieren, welche Tiefe beim Textverstehen aufgrund formaler Beschränkungen in den Repräsentationssprachen überhaupt erreichbar ist.

Logik anbieten. Im Gegensatz dazu formulieren Artale & Franconi (1994) eine temporale Erweiterung zu einer terminologischen Logik und untersuchen deren formale Eigenschaften bezüglich Konzeptklassifikation und Instanzmatching. Ebenso wie Schmiedel (1990) gehen sie von einer total geordneten Menge von Zeitpunkten aus, um die formale Semantik der zusätzlichen temporalen Quantoren zu definieren. Einen Schwerpunkt bei Konzeptbeschreibungen mit temporaler Struktur bilden komplexere Handlungen (ACTIONS), die aus einer zeitlich partiell geordneten Sequenz von Subhandlungen bestehen (Artale & Franconi, 1998).³

Im Textverstehenssystem SYNDIKATE finden Erweiterungen zu terminologischen Sprachen um temporale Quantoren keine Anwendung. Vielmehr sind zeitliche Bezüge (wie auch spatiale Relationen) durch expressive Mittel mit einer schwachen Semantik repräsentiert.⁴ Dazu ist ein Inventar an Relationen definiert, die explizit im Relationsbezeichner temporale Symbole – wie beispielsweise PRECEDING-EVENT, INITIAL-TIME-POINT usw. – verwenden. Jedoch verfügen diese Relationsbezeichner gegenüber anderen (z.B. AGENT oder HAS-HARD-DISK) über keinen besonderen epistemologischen Status. Demzufolge gibt es kein auf terminologischer Logik basierendes Folgern über die temporale Struktur von Ereignissen oder die Gültigkeit von Aussagen relativ zu Referenzzeitpunkten. Dennoch repräsentiert die während der Textanalyse gewonnene inhaltliche Interpretation temporale Bezüge durch eine differenzierte Menge an Relationen. Zur Illustration sei der Beispielsatz 12 eingeführt:

12. Nach der Installation der Festplatte verfügt der Computer über 20.3 GB Speicher.

Die Präposition *“nach”* stellt eine temporale Beziehung zwischen Zustand *“verfügt”* (POSSESS.3-06) und der Handlung *“Installation”* (INSTALL.1-03) her. Der semantische Interpretationsprozeß der Präpositionalphrase berechnet die konzeptuelle Relation

³Die Möglichkeit, eine (zumindest partielle) Ordnung für Konzepte in Handlungssequenzen zu definieren, wäre in SYNDIKATE für die in Kapitel 5.2.1.2 eingeführten Konzepte des Typs VERB-MODEL von Bedeutung.

⁴Wegen der starken Analogie zwischen temporalem und spatialem Schlußfolgern befaßt sich dieses Kapitel nur mit dem temporalen Folgern. Entsprechende Erweiterungen terminologischer Logik, die dediziertes spatiales Folgern ermöglichen, finden sich beispielsweise in Haarslev et al. (1998). Einen Vorschlag zur Modellierung spatialer Beziehungen mit *ALC* bieten Schulz et al. (2000).

PRECEDING-EVENT wie in Abb. 5.1 zu sehen ist.⁵ Darüber hinaus werden den Verb-Konzepten die an den finiten Verbformen im Satz – also auch den Modal- und Auxiliärverben – anzutreffenden Merkmale für Modalität, Tempus und Genus in Form einer entsprechenden Marke übertragen. Auch diese Form der Interpretation macht die im Satz durch die Flexion des finiten Verbs markierte Information auf der inhaltlichen Repräsentationsschicht verfügbar, die Semantik dieser Markierung bleibt indessen ebenfalls unterdeterminiert.

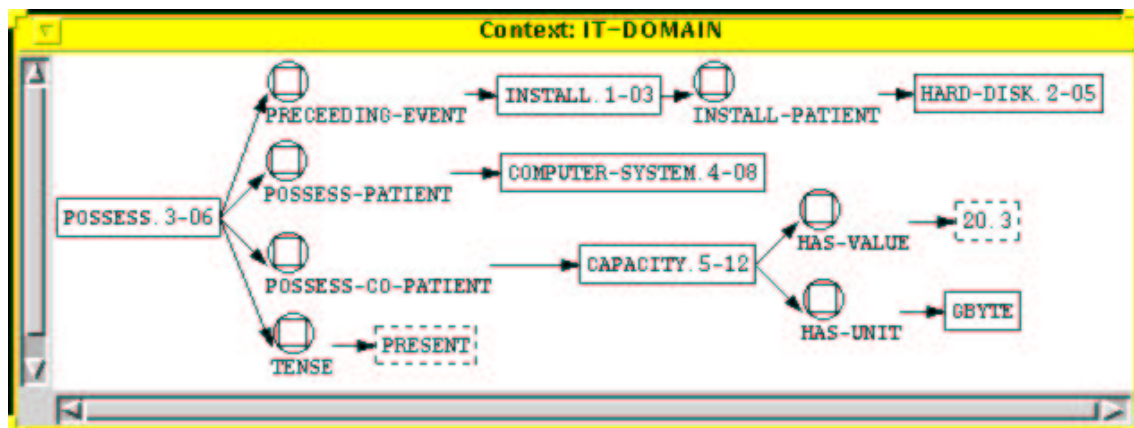


Abbildung 5.1: Konzeptuelle Repräsentation temporaler Bezüge

Die damit erzielte Tiefe des Verstehens reduziert weiter interpretierbare temporale Bezüge auf einfache Relationsbeziehungen. Darüber hinaus offenbart ein genauerer Blick auf den Beispielsatz, daß über den temporalen Bezug eine Bedingung kodiert ist, d.h., erst wenn die Installation erfolgt, steht der Festplattenspeicher auch tatsächlich zur Verfügung. Dieser Verlust an Information legitimiert sich durch zwei Argumente. Zum einen steht in SYNDIKATE die Breite der Verstehensleistung im Vordergrund, so daß auf eine Interpretation aller impliziten Bezüge zugunsten einer breiteren Abdeckung verzichtet wird. Zum anderen – und dieses Argument besitzt im Anwendungskontext ebenfalls Gewicht – implementiert kein auf terminologischer Logik basierendes Wissensrepräsentationssystem temporales Reasoning, auch wenn die dazugehörigen Sprachen und Klassifikationsalgorith-

⁵In der Abb. 5.1 ist nicht verzeichnet, daß eine Regel der konzeptuellen Interpretation auch COMPUTER-SYSTEM.4-08 mit CAPACITY.5-12 über HAS-HARD-DISK HAS-DISK-CAPACITY relationiert. Die Referenzidentität der im berechneten Relationspfad implizit enthaltenen Festplatte mit HARD-DISK.2-05 entdeckt SYNDIKATE beim gegenwärtigen Stand der Implementation nicht.

men beschrieben sind (Artale & Franconi, 1998).

5.1.4 Quantifikation und Beschreibungslogiken

Im Rahmen einer formal-linguistischen Betrachtung von Satz- wie auch Textanalysen besteht das wahrscheinlich wichtigste Problem darin, die unterschiedlichen Ebenen der natürlichsprachlichen Quantifikation einschließlich der möglichen Skopusambiguitäten formal zu repräsentieren. Mit der Quantifikation sind eine ganze Reihe von Aspekten der semantischen Interpretation verbunden, die einen essentiellen Beitrag zum inhaltlichen Verständnis von Äußerungen leisten. Aus der Vielzahl der in der Literatur diskutierten Fälle seien hier beispielhaft einige herausgegriffen, ohne den Anspruch auf eine vollständige Übersicht zu erheben (vgl. hierzu Allgayer & Reddig (1990)):

- Bei komplexen Nominalphrasen sind deren Denotate unter Berücksichtigung des syntaktischen Kontexts zu ermitteln (Bsp.: “Der Staat sorgt sich um Frauen mit Kindern im Gefängnis”). Daher kann jede Nominalphrase als quantifiziert verstanden werden.
- Im semantischen Interpretationsprozeß ist zu bestimmen, ob Aussagen in einem Satz generische Eigenschaften eines Konzepts, prototypische Eigenschaften eines Konzepts oder Aussagen über ein Individuum bzw. Mengen von Individuen enthalten.
- Auflösung von Lesartalternativen, wenn quantifizierte Nominalphrasen in Zusammenhang mit Verbalphrasen zu interpretieren sind (kollektive, distributive und kumulative Lesart). Diese Form der semantischen Interpretation macht stark von enzyklopädischem Wissen Gebrauch (vgl. den vielstrapazierten Satz “Vier Frauen spielen Klavier”).
- Mengenbeziehungen zwischen Diskursreferenten in der textuellen Progression (Bsp.: “*Fünf Kinder* sitzen im Sandkasten. *Die beiden Mädchen* ...”) bzw. auch Subkonzeptbeziehungen zwischen Diskursreferenten (Bsp.: “*Das Floptical-Laufwerk* kommt in **drei Versionen**. Das *externe Laufwerk für den Mac*, das *externe Laufwerk für den PC* und ein *internes Laufwerk*.”)
- Schließlich spielen diejenigen quantifizierenden natürlichsprachlichen Ausdrücke eine wichtige Rolle, die mit den Quantoren von logischen Sprachen erster Ordnung nicht repräsentierbar sind. Diese Ausdrücke werden beispielsweise in der *Generalized*

Quantifier Theory (Barwise, 1987; Barwise & Cooper, 1981) aufgegriffen (Bsp.: viele, wenige, höchstens etc.).

Gemessen an den weitreichenden Ansprüchen an eine Wissensrepräsentationssprache, wie sie sich aus der vorstehenden skizzenhaften Klassifikation unterschiedlicher Phänomenklassen ergeben, nehmen sich die Ausdrucksmöglichkeiten der in der Diskussion als Referenzsprache verwendeten terminologischen Sprache $\mathcal{ALC}$ zunächst überaus bescheiden aus. Interpretiert man terminologische Axiome prädikatenlogisch als Implikationen bzw. Äquivalenzaussagen, dann verfügt $\mathcal{ALC}$ lediglich über eingeschränkte Mittel, um all- bzw. existenzquantifizierte Axiome darzustellen.⁶

Zunächst beschreiben einfache Konzeptdefinitionen Eigenschaften für Klassen von Individuen. Dadurch sind diese implizit allquantifiziert. Diese Allquantifizierung wird bei der Übertragung terminologischer Logik in Prädikatenlogik offensichtlich. Zum Beispiel ließe sich eine primitive Konzeptdefinition⁷

$$\text{COMPUTER-SYSTEM} \sqsubseteq \text{HARDWARE}$$

folgendermaßen in Prädikatenlogik übertragen:

$$\forall(x) : \text{Computer-System}(x) \Rightarrow \text{Hardware}(x).^8$$

Neben dieser impliziten Form der Allquantifizierung, existieren – wie bei der Referenzsprache $\mathcal{ALC}$ – in der Mehrzahl der terminologischen Sprachen der Existenz- und der Allquantor als konzeptformende Operatoren auf der rechten Seite eines terminologischen Axioms, um Typrestriktionen für konzeptuelle Rollen zu fassen. Zur Veranschaulichung mag das bereits verwendete Beispiel eines Computers dienen, der mindestens eine CPU hat.

⁶Ein Vergleich zwischen terminologischen Logiken und Prädikatenlogik findet sich in Borgida (1996).

⁷Bei der Übertragung wird der Operator für primitive Konzeptdefinitionen “ $\sqsubseteq$ ” in die Implikation “ $\Rightarrow$ ” übersetzt, wohingegen der Operator für definitorische Konzeptdefinitionen “ $\doteq$ ” zu “ $\Leftrightarrow$ ” übersetzt wird.

⁸Bei dieser vereinfachenden Darstellung werden Besonderheiten terminologischer Systeme wie etwa Vererbung außer acht gelassen.

$$\text{COMPUTER-SYSTEM} \doteq \text{HARDWARE} \sqcap \exists \text{HAS-PART. CPU}$$

Eine Übertragung in Prädikatenlogik führt zum folgenden Axiom:

$$\forall(x) : \text{Computer-System}(x) \Leftrightarrow \text{Hardware}(x) \wedge \exists(y) : (\text{CPU}(y) \wedge \text{has-part}(x, y))$$

In $\mathcal{ALC}$ nicht formulierbar sind terminologische Axiome, die in ihrer prädikatenlogischen Übersetzung einen Existenzquantor im Antezedenten einer Implikation aufweisen. Auch hybride Spracherweiterungen um eine ausdrucksmächtigere assertionale Ebene, die T-Box-Ausdrücke auf der rechten Seite zulassen, schaffen hier keine Abhilfe, da assertionale Axiome Aussagen über konkrete Individuen machen, die sowohl in terminologischer Logik als auch in Prädikatenlogik in Form von (Individuen-)Konstanten notiert werden.

Um den am Beispiel der Quantifikation deutlich zutage tretenden Expressivitäts-Defiziten zu begegnen, gibt es zu diesem Problem eine Reihe von Arbeiten über terminologische Logiken. Sie verwenden Spracherweiterungen im Sinne einer starken Semantik, die die Menge von Operatoren und epistemischen Primitiven sowie die deduktive Reasoningkomponente substantiell erweitern.

Als Ausgangspunkt für eine Erweiterung terminologischer Logik um eine an natürlicher Sprache orientierte Quantifikation dient Quantz (1992) die Generalized Quantifier Theory (Barwise & Cooper, 1981). Dieser zufolge bezeichnen generalisierte Quantoren, wie sie in natürlichsprachlichen Äußerungen verwendet werden, binäre Quantoren, die Beziehungen zwischen Tupeln von Teilmengen einer Domäne Δ ausdrücken. In der terminologischen Sprache wird dieser Tatsache dadurch Rechnung getragen, daß zusätzlich eingeführte binäre Quantoren Beziehungen zwischen Mengen auf der Grundlage einer modelltheoretischen Semantik definieren. Neben Konzeptbezeichnern sind in diesen Ausdrücken auch Objektbezeichner zulässig, die numerisch definierte Mengen repräsentieren. Durch diese Spracherweiterung lassen sich in etwa Sätze wie “Mehr als 3 Hersteller sind pleite” repräsentieren: $\geq 3(\text{Producer}, \text{Bankrupt})$. Sowohl *Producer* als auch *Bankrupt* sind Teilmengen einer Domäne Δ , d.h. $\text{Producer}, \text{Bankrupt} \in 2^\Delta$. Das Axiom drückt demnach aus, daß die Schnittmenge aus beiden mindestens drei Elemente enthält.

Franconi (1993) erweitert die terminologische Konzeptsprache $\mathcal{ALC}$ um Pluralquantifikatoren zur Sprache $\mathcal{ALCS}$, indem er Mengen von Individuen mit Hilfe einer ausgezeichneten *has-element*-Relation zu Kollektionen zusammenfaßt. Individuen können dabei

unterschiedlichen Kollektionen angehören. Die Zugehörigkeit eines Elements zu einer Kollektion wird in Form eines Terms und nicht über ein Prädikat ausgedrückt. Aufbauend auf Mengenbeziehungen zwischen Kollektionen läßt sich durch die Verwendung der Pluralquantifikatoren ausdrücken, ob Handlungen und Events kollektiv, distributiv oder kumulativ zu lesen sind. Die distributive Lesart und die kollektive Lesart sind jeweils ein Spezialfall der kumulativen Lesart. D.h., im ersten Fall ist die Anzahl der zu bildenden Unterkollektionen gleich der Anzahl der Elemente einer Kollektion, wohingegen im zweiten Fall nur eine einzige Unterkollektion zu bilden ist, nämlich die, die alle Elemente enthält. Bei der Analyse natürlicher Sprache wird daher jede Interpretation, die eine Kollektion involviert zunächst auf die kumulative Lesart abgebildet (Franconi, 1994). Zu einem späteren Zeitpunkt in einem Text hinzukommende Information dient gegebenenfalls der Disambiguierung bzw. der Spezialisierung der Lesart.

Auch bei Allgayer & Reddig (1990) bildet die Generalized Quantifier Theory den Ausgangspunkt zu den Überlegungen, wie mit terminologischer Logik quantifizierende natürlichsprachliche Phrasen darzustellen sind. Aufbauend auf der hybriden Wissensrepräsentationssprache SB-ONE modellieren sie zunächst Mengenbeziehungen mit den Konzeptdefinitionen aus der terminologischen Sprache, die SB-ONE zur Verfügung stellt (im Sinne einer schwachen Semantik). Jedoch gelangen Allgayer & Reddig (1990) zu der Auffassung, daß Mengenbeziehungen adäquat nur durch eine Erweiterung der Beschreibungssprache auf der epistemischen Ebene zu fassen sind. Die hierzu notwendigen Erweiterungen sind in der Sprache SB-ONE+ gefaßt (Allgayer, 1990).

Die terminologische Beschreibungssprache in SB-ONE+ wird dahingehend erweitert, daß für den Typen eines Füllers zusätzlich anzugeben ist, ob es sich um Individuen oder um Mengen handelt.⁹ Darüber hinaus sind Relationen entweder Mengen-Relationen oder Element-Relationen, je nachdem ob die Domain eine Menge oder ein Individuum verlangt. Beispielsweise bezeichnet HAS-AGE eine Element-Relation, weil in der Domäne PERSON definiert ist, wohingegen AVERAGE-AGE eine Mengen-Relation darstellt, weil der Wertebereich der Relation eine Aussage über eine Menge an Individuen beschreibt. Entsprechend enthält die A-Box die Differenzierung zwischen Individuen und Mengen. Dadurch

⁹Die formale Semantik muß bei Konzeptdefinitionen entsprechend differenziert werden. Bei der Festlegung auf Individuen eines Konzepttyps erfolgt die Abbildung im Wertebereich gegeben eine Domäne Δ wie gehabt nach 2^Δ , bei der Festlegung auf Mengen von Individuen eines Konzepttyps hingegen nach 2^{2^Δ} , d.h. Mengen von Mengen aus der Domäne Δ .

können Instanzen zu Mengen zusammengefaßt und assertionale Axiome zu beiden A-Box-Konstrukten aufgestellt werden.

Einen Schritt weiter als SB-ONE+ geht die Sprache *SBTWO*, da sie zusätzlich zu den Erweiterungen von SB-ONE+ eine Menge an deskriptiven Operatoren einführt. Diese deskriptiven Operatoren dienen dazu, über die durch SB-ONE+ abgedeckten Mengenbezüge hinaus weitere natürlichsprachliche Phänomene formal in Beschreibungslogik einzuführen. In diesen Bereich gehören bestimmte und unbestimmte Nominalphrasen über Mengenterme, prototypische Aussagen sowie Konzeptdefinitionen mit Referenzen auf A-Box-Ausdrücke.

Abschließend betrachtet gehen die drei in diesem Kapitel vorgestellten Ansätze, mit denen natürlichsprachliche Quantifikationsphänomene in terminologischer Logik zu repräsentieren sind, in dieselbe Richtung. Auf der deskriptiven Ebene führen sie neue Sprachkonstrukte ein, die die Repräsentation von Mengen, aber auch von quantifizierenden Nominalphrasen in Beschreibungslogik ermöglichen.¹⁰ Jedoch bleibt festzuhalten, daß keines der augenblicklich verfügbaren Wissensrepräsentationssysteme die formal beschriebenen Erweiterungen terminologischer Logik implementiert. Damit erweist sich aus einer anwendungsbezogenen Perspektive betrachtet die mangelnde Expressivität auf der Ebene der Quantifikation als ein Manko KL-ONE-artiger Systeme. Im übrigen gelten für terminologische Sprachen mit Mengenkonstrukten und erweiterter Quantifikation dieselben Aussagen wie für alle anderen Fälle, in denen die Expressivität erhöht wird. Aufgrund der Wechselwirkung zwischen Expressivität auf der einen Seite und Berechenbarkeit von Subsumption und Vollständigkeit der Klassifikationsalgorithmen auf der anderen Seite gehen in terminologischen Systemen Spracherweiterungen mit Einschränkungen an anderer Stelle einher.

Welche Implikationen ergeben sich aus den Ausführungen zu terminologischen Logiken für die im Textverstehenssystem SYNDIKATE erzielbare Verstehenstiefe? Zunächst verzichtet SYNDIKATE auf die in Texten enthaltende Differenzierung, ob Aussagen generisch, prototypisch oder auf ein Individuum bezogen sind, weil für jedes Diskursobjekt grundsätzlich eine Instanz generiert wird. Temporale wie auch spatiale Beziehungen zwischen Diskursobjekten sind durch *ALC*-konforme Konzept- und Relationsbeschreibungen repräsentiert. Die im Textanalyseprozeß gewonnene Inhaltsrepräsentation enthält nahezu

¹⁰Ebenfalls integrierte Erweiterungen im automatischen deduktiven Reasoning wurden hier nicht thematisiert.

keine Information über Quantifikation.

Allen diesen Einschränkungen zum Trotz können die erzielten Inhaltsrepräsentationen ggf. als Aufsatzpunkt für weitere Interpretationen dienen. D.h., augenblicklich implementiert SYNDiKATE einen ersten Interpretationsschritt, der – unter den in realistischen Anwendungsumgebungen gegebenen Beschränkungen – versucht, die wesentlichen Aussagen zu extrahieren. Insgesamt geht die Verstehensleistung über die im MUC-Kontext geforderte hinaus, denn sie bezieht komplexere Relationierungsmuster zwischen Diskursobjekten mit ein, prüft bei semantischen Konsistenzchecks nicht nur die reine Argumentstruktur und berücksichtigt schließlich textuelle Phänomene bei der Wissensextraktion.

Abschließend sei angemerkt, daß der Schwerpunkt beim Textverstehen sich auch durch die Bedürfnisse derjenigen Nutzer bestimmt, die im Anschluß das in den Textwissensbasen abgelegte Wissen verwenden. In diesem Sinne orientiert sich SYNDiKATE an einem sehr allgemeinen Nutzermodell.

5.1.5 Wissensrepräsentationssysteme für Beschreibungslogiken

Neben den formalen Betrachtungen, die sich mit der Expressivität terminologischer Sprachen und deren theoretischen Berechnungseigenschaften befassen, sind auch die Laufzeiteigenschaften der Wissensrepräsentationssysteme von Interesse, weil diese eine große Anzahl an Konzept- und Relationsdefinitionen zu verarbeiten haben. Schließlich dienen die von ihnen implementierten Sprachen als Grundlage der Wissensmodellierung. Daher befassen sich Heinsohn et al. (1994) in einem empirischen Vergleich von KL-ONE-artigen Wissensrepräsentationssystemen – die hier im einzelnen nicht aufgelistet werden – mit dem tatsächlichen Laufzeitverhalten in einer Reihe experimenteller Tests.

Diese Tests beziehen sich insbesondere auf die Vollständigkeit der Klassifikationsalgorithmen (qualitativer Test) sowie den zeitlichen Klassifikationsaufwand relativ zur Anzahl der zu klassifizierenden Konzepte (quantitativer Test). Im Vergleich zu den in die empirische Untersuchung mit einbezogenen sechs Wissensrepräsentationssystemen schneidet LOOM insgesamt am besten ab, da es die größte Ausdrucksmächtigkeit besitzt und in den verschiedenen Laufzeittests mit knappem Abstand zum schnellsten System auf Rang zwei platziert ist.¹¹

¹¹Interessanterweise verfügt zum Zeitpunkt der Untersuchung von Heinsohn et al. (1994) keine der in terminologischer Logik implementierten Wissensbasen über mehr als 500 Konzepte und Relationen. Die

Eine weitere empirische Arbeit über das Laufzeitverhalten von terminologischen Systemen stammt von Horrocks (1995). Allerdings fokussiert diese Untersuchung auf LOOM und GRAIL sowie auf die Integration transitiver Rollen in die Konzeptbeschreibungssprache. Darüber hinaus wird seit einiger Zeit an der Entwicklung eines objektiven Benchmark-Tests für Wissensrepräsentationssysteme gearbeitet (Horrocks & Patel-Schneider, 1998).

5.1.6 Seitenblick auf weitere Wissensrepräsentationssprachen

Verallgemeinernd kann festgestellt werden, daß größere Ontologien, sofern sie in einem Repräsentationsformalismus mit wohldefinierter Semantik konzipiert sind, entweder direkt auf KL-ONE-artige Systeme zurückgreifen oder aber sich zumindest an diesen orientieren (z.B. das GUM bzw. CYKL, vgl. hierzu Kapitel 5.3.1). Interessanterweise gilt dies mittlerweile auch in der medizinischen Domäne (z.B. K-REP, GRAIL oder neuerdings SNOMED-RT vgl. hierzu Kapitel 5.3.2). Andere formal fundierte Wissensrepräsentationsformalismen spielen im Bereich des Ontological Engineering keine besondere Rolle mehr, obwohl sie prinzipiell dazu geeignet wären. Zu dieser Gruppe gehören beispielsweise die etwa zeitgleich zu den KL-ONE-artigen Systemen entstandenen *Conceptual Graphs* (Sowa, 1984) oder SNEPS (Shapiro, 1979).

Conceptual Graphs rekurren auf ein formales logisches System, das dementsprechend über eine formale Semantik verfügt. Diese Wissensrepräsentationssprache stellt formalsprachliche Konstrukte bereit, die in besonderem Maße geeignet sind, natürlichsprachliche Ausdrücke abzubilden (Sowa, 1991), indem beispielsweise Phänomene der Quantifikation, Skopus von Propositionen oder auch temporale Aussagen repräsentiert werden können. Insbesondere bei Systemen zur Textanalyse in der Medizin dominieren die konzeptuellen Graphen als Quasi-Standard (aus den in Kapitel 2.2.1 angeführten Beispielen seien nochmals herausgegriffen Baud et al. (1995), Volot et al. (1994), Bernauer (1992) und Friedman et al. (1994)).¹² Im Gegensatz dazu existiert keine größere Ontologie, die unter Verwendung konzeptueller Graphen implementiert ist.

Ähnliche Aussagen lassen sich auch über SNEPS treffen. Als Wissensrepräsentationssprache ermöglicht SNEPS dem Knowledge Engineer die eigenständige Definition von epi-

GRAIL- wie auch die SYNDIKATE-Wissensbasis haben diese Grenze um ein Vielfaches überschritten.

¹²Die Beziehung zwischen konzeptuellen Graphen und Beschreibungslogiken wird beschrieben in Biebow & Chaty (1993) und Ingenerf (1996).

stemischen Primitiven (Shapiro & Rapaport, 1992), wodurch die Semantik der Sprache unscharf wird. Andererseits lassen sich durch diese Mechanismen bei der Repräsentation natürlichsprachlicher Äußerungen Bezüge zwischen Propositionen flexibel darstellen. Folglich kommt SNEPS durchaus als Zielsprache zur Anwendung, wenn natürliche Sprache zu repräsentieren ist (Ali & Shapiro, 1993), zur Modellierung mächtiger Ontologien scheidet SNEPS als Wissensrepräsentationssprache indessen aus.

Vor dem Hintergrund dieses grob skizzierten Zusammenhangs wird der Vorteil des in SYNDIKATE verwendeten beschreibungslogischen Ansatzes deutlich. Einerseits dient mit LOOM ein auf Beschreibungslogik basierendes KL-ONE-artiges System der Repräsentation von ontologischem Wissen. Andererseits fungiert dieses System zugleich als Zielsprache des Textverstehensprozesses. Dies führt zu einer erheblichen Reduktion des Modellierungsaufwands, da das ontologische Wissen unmittelbar in die Textanalyse eingebracht werden kann. Darüber hinaus vermeidet die Architektur von SYNDIKATE unplausible Kombinationen von Wissensrepräsentationsformalismen wie etwa bei (Rassinoux et al., 1998), die GRAIL für das Ontological Engineering verwenden und konzeptuelle Graphen als Zielsprache der Textanalyse, wodurch ein eigentlich überflüssiges Übersetzungsmodul erforderlich wird.

5.2 Ontological Engineering im Kontext der Wissensextraktion aus Texten

Jede Form von *tieferem* Textverstehen basiert auf a priori bereitgestelltem Weltwissen und Schlußfolgerungswissen, das auf diesen Weltwissensstrukturen operiert. Damit gewinnt das Problem des Ontological Engineering für die Wissensextraktion aus Texten an Gewicht. In SYNDIKATE kommt dem Weltwissen, wie bereits erwähnt, sogar eine doppelte Funktion zu. Einerseits unterstützt es die dem Verstehen unterlegten Reasoningprozesse über konzeptuelle Wohlgeformtheitsbedingungen (das semantische Filtern). Andererseits dient es zugleich als Zielsprache, die das aus den Texten extrahierte Wissen repräsentiert.

Zentral für das Knowledge Engineering ist die Grundidee, Probleme der verschiedenen Anwendungsschichten, die auf dem zur Verfügung stehenden Wissen aufsetzen, nicht in das Domänenmodell des fraglichen Weltausschnitts hineinzutragen. Innerhalb der Medizinischen Informatik wurde dieser Gedanke beispielsweise durch das Konzept des *Ter-*

minology Servers aufgebracht (Rector et al. (1995b)), dessen Kernsystem die semantische Interoperabilität zwischen einer breiten Palette unterschiedlichster Anwendungen gewährleisten soll. Vor diesem Hintergrund beleuchtet das Kap. 5.2 die Modellierungsprinzipien des Ontological Engineering in SYNDIKATE.

5.2.1 Konzeption und Legitimation des ontologischen Entwurfs in SYNDIKATE

Zwar stellt die von SYNDIKATE geleistete semantische Interpretation, die überdies noch von der konzeptuellen Interpretation flankiert wird, auf der Ebene der Verstehensstufe eine Erweiterung zu den Informationsextraktionssystemen dar. Dennoch muß erwähnt werden, daß die Verstehensleistung durch die Festlegung auf eine Wissensrepräsentationssprache einen begrenzenden Faktor hat (vgl. hierzu die Ausführungen zu den Komplexitätseigenschaften terminologischer Logiken in Kap. 5.1).

Vor dem Hintergrund der Entscheidung, als formale Beschreibungssprache eine terminologische Logik zu verwenden, um sowohl allgemeines als auch domänenspezifisches konzeptuelles Wissen zu repräsentieren, ist die benötigte Ausdrucksmächtigkeit des verwendeten formalen Systems aus den konkreten Anforderungen des Textverstehens wie auch der ontologischen Struktur des Diskursbereichs zu bestimmen.

Im Kontext des Textverstehenssystems SYNDIKATE lassen sich für das domänenunabhängige UPPER-MODEL wie für die beiden Spezialwissensbasen IT und MED Minimalanforderungen an eine Beschreibungslogik definieren. Zu den in unserem Ansatz zum Textverstehen unverzichtbaren konzeptformenden Operatoren gehören (vgl. hierzu 3.2):

- die *Konjunktion* von Konzepten zur Generierung multipler Vererbung in Konzepthierarchien. Zum Beispiel wird das Konzept GASTRITIS sowohl von INFLAMMATION als auch von GASTRIC-DISEASE subsumiert. Dadurch erbt GASTRITIS von einem der Elternkonzepte die ätiologischen Eigenschaften und vom anderen die Lokalisation. Multiple Hierarchien erlauben demzufolge unterschiedliche „Sichten“ auf ein und dasselbe Konzept.
- die *Disjunktion* (Veroderung) von Konzepten, da z.B. in der Medizin-KB der Domainbereich der Relation POPULATED-BY-MICROORGANISMS durch BODY-SUBSTANCE oder ANATOMICAL-STRUCTURE zu definieren ist. Ein ähnlich gelagerter Fall aus

der Informationstechnologie besteht aus dem Füller der Relation INSTALL-PATIENT des Verbkonzepts INSTALL. Der Wertebereich dieser Relation ist die Disjunktion der Konzepte HARDWARE und SOFTWARE.

- Die *Negation* spielt eine wichtige Rolle u.a. bei der Auflösung referentieller Phänomene, aber auch bei der Disambiguierung. Durch die textuelle Progression bedingte Erweiterungen der Beschreibung von Diskursentitäten führen zu Inkonsistenzen im entsprechenden Kontext der Textwissensbasis, wenn eine Instanz auf zwei Konzepte abgebildet wird, die mengentheoretisch kein gemeinsames Element in der Extension besitzen (was durch die Negation ausgedrückt wird). D.h., diese Inkonsistenz einer Instanz dient als Indikator dafür, daß die betroffene Interpretation nicht weiter zu verfolgen ist (vgl. Kapitel 4.3).¹³
- der *Allquantor* und der *Existenzquantor* für Typbeschränkungen bei Rollendefinitionen. Ein COMPUTER-SYSTEM ist u.a. dadurch definiert, daß es eine Rolle HAS-HARD-DISK gibt. *Jeder* dieser Füller muß vom Typ HARD-DISK sein, unabhängig von der Anzahl der möglichen Füller. Im Gegensatz dazu beschreibt die Rolle DEVELOPS das Konzept PRODUCER dadurch, daß mindestens ein Füller vom Typ PRODUCT existiert. Damit sind ausdrücklich auch andere Rollenfüller erlaubt.
- *kardinale Beschränkungen* für die Anzahl von Rollenfüllern eines Konzepts. Eine (gesunde) Hand beispielsweise verfügt über *genau* fünf Finger (eine ausführliche Begründung für die Notwendigkeit von Kardinalitätsbeschränkungen in der Medizin liefern Smart & Roux (1995)).¹⁴ Zum Spezialwissen der IT gehört zum Beispiel, daß an einen SCSI-Hostadapter *maximal* sieben Devices angeschlossen werden können. Auf der Ebene terminologischer Axiome muß demzufolge eine entsprechende Wertebeschränkung in die Konzeptdefinition eingefügt werden.

¹³Wegen des besonderen Stellenwerts der Disambiguierung beim Textinterpretationsprozeß findet sich im oberen Bereich der Konzepthierarchie von SYNDIKATE eine weitgehende Partitionierung, d.h., die in den Teilkonzeptgraphen der Partition enthaltenen Konzepte haben disjunkte Extensionen (vgl. wiederum 3.2).

¹⁴Das innerhalb der Medizin nicht unerhebliche Problem, neben dem gesunden auch den pathologischen Organismus zu modellieren, kann hier nicht thematisiert werden. Jedoch sind beispielsweise numerische Restriktionen im Fall eines erkrankten oder atypischen Organismus nicht eindeutig festlegbar. Obwohl ein Mensch normalerweise über zwei Nieren verfügt, können es auch drei oder nur eine sein. Aus diesem Sachverhalt ließe sich ein Argument für das Verwenden von default-logischen Spracherweiterungen machen.

Eine weitere in dieser Aufzählung nicht angeführte – allerdings SYNDIKATE-spezifische – Anforderung besteht darin, die Textwissensbasis während der Analyse eines Eingabetexts inkrementell um Axiome über die eingeführten Diskursentitäten zu erweitern. Im Kapitel 3.2.5 zum Textverstehen wurde aufgezeigt, daß die Analyse von Einzelsätzen wie auch die Resolution anaphorischer Phänomene zu einer kontinuierlichen Akkumulation von konzeptuellen Relationierungen an Diskursentitäten führt. Bedingt durch die Auswahl unserer Textkorpora finden sich in den medizinischen Pathologieberichten vorwiegend Aussagen zu Individuen (Bsp.: *“das Partikel”*, welches sich unter dem Mikroskop befindet), wohingegen Texte aus der Informationstechnologie hauptsächlich Konzepte beschreiben (Bsp.: *“Sony’s Vaio kommt ausschließlich mit TFT-Displays”*).¹⁵ Weil das in SYNDIKATE verwendete Wissensrepräsentationssystem LOOM nur die inkrementelle Erweiterung der ABox, nicht aber die von Konzeptbeschreibungen auf der TBox-Ebene zuläßt, erfolgt der Aufbau der Textwissensbasen allein in Form von assertionalen Axiomen. Damit geht die Unterscheidung zwischen generischen Aussagen und Aussagen über Individuen zunächst verloren (vgl. Azzam (1994)). Das Textverstehen hört jedoch nicht beim kontinuierlichen Erweitern einer Textwissensbasis durch assertionale Axiome auf, sondern aufbauend auf den assertionalen Erweiterungen in der Textwissensbasis zieht die deduktive Komponente weitergehende Schlußfolgerungen über die Konzeptzugehörigkeit von Instanzen.

Diese Schlußfolgerungen sind für das faktenorientierte Retrieval essentiell. Zur Illustration sei der Beispielsatz 13 gegeben und zu diesem ein kleines Fragment aus der Konzepthierarchie in der IT-Domäne eingeführt (Abb. 5.2). Die Konzeptdefinitionen sind hier in der für LOOM üblichen Konzeptdefinitionssprache wiedergegeben. Der konzeptformende Operator *the* ist eine syntaktische Kontraktion aus $\forall R : C$ sowie $\exists_{\geq n} R$ und $\exists_{\leq n} R$ mit $n = 1$, d.h., es gibt genau einen Füller der Relation R mit Typ C . Die Konzepte HD-FLOPPY-DISK und 3.5-INCH-FLOPPY-DISK unterscheiden sich in den zulässigen Wertemengen der Relationen für Durchmesser und Kapazität. HD-FLOPPY-DISKS gibt es sowohl mit 3.5 als auch mit 5.25 Zoll Durchmesser, wohingegen 3.5-INCH-FLOPPY-DISKS mit zwei unterschiedlichen Kapazitäten ausgeliefert werden.

13. eine 3,5-Zoll-Diskette mit 1,44 MB Kapazität.

Da der Satz 13 zwei Inhaltswörter aufweist, finden sich in der Textwissensbasis zwei In-

¹⁵Sehr selten finden sich in der Informationstechnologie auch Referenzen auf Individuen (Bsp.: *“Das Gerät im Test”*).

```
(defconcept HD-Floppy-Disk
  :is (:and Floppy-Disk
    (:the has-capacity
      (:and Capacity
        (:the has-unit (:one-of MByte))
        (:the has-value (:one-of 1.44 1.2))))
    (:the has-diameter
      (:and Diameter
        (:the has-unit (:one-of Inch))
        (:the has-value (:one-of 3.5 5.25)))))))

(defconcept 3.5-Inch-Floppy-Disk
  :is (:and Floppy-Disk
    (:the has-capacity
      (:and Capacity
        (:the has-unit (:one-of MByte))
        (:the has-value (:one-of 1.44 0.72))))
    (:the has-diameter
      (:and Diameter
        (:the has-unit (:one-of Inch))
        (:the has-value (:one-of 3.5)))))))
```

Abbildung 5.2: Konzeptdefinition für Floppy-Disketten

stanzen, FLOPPY.1-01 und CAPACITY.2-02. Darüber hinaus binden die Relationen HAS-VALUE und HAS-UNIT die numerische Angabe “1.44” und die Einheit “MB” an die Instanz CAPACITY.2-02, welche die Speicherkapazität repräsentiert. Sobald die Präposition “mit” ihren syntaktischen Kopf sucht, wird die entsprechende semantische Interpretationsregel instantiiert. Als Ergebnis des Interpretationsprozesses erfolgt eine Relationierung von FLOPPY.1-01 mit CAPACITY.2-02 über HAS-CAPACITY. Auf der Grundlage der terminologischen Axiome in Abb. 5.2 und der im Verlauf der Satzanalyse in der Textwissenbasis hinzugefügten assertionalen Axiome leitet die deduktive Komponente des Wissensrepräsentationssystems ab, daß die Instanz FLOPPY.1-01 auch zur Extension des Konzepts HD-

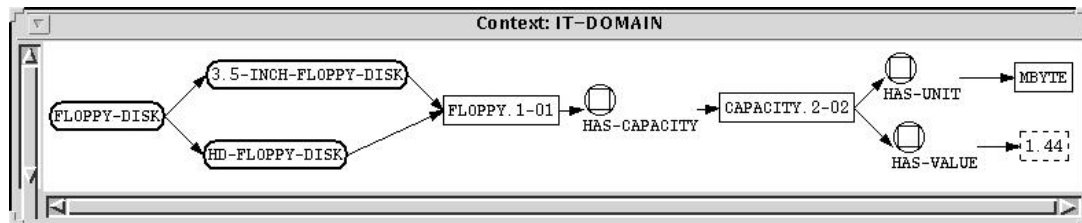


Abbildung 5.3: Reasoning mit Instanzen

FLOPPY-DISK gehört.

Mit diesem Beispiel läßt sich verdeutlichen, daß die inkrementelle Erweiterung von Diskursentitäten durch den Interpretationsprozeß nicht nur zu einer Abbildung von Instanzen auf speziellere Konzepte führt, sondern auch zu einer Abbildung auf Konzepte im Konzeptgraphen, die in keiner Subsumptionsbeziehung zueinander stehen. In Abb. 5.3 findet sich eine graphische Darstellung des beschriebenen Sachverhalts. FLOPPY.1-01 steht sowohl zu HD-FLOPPY-DISK als auch zu 3.5-INCH-FLOPPY-DISK, also zwei Geschwisterkonzepten, in einer INSTANCE-OF Beziehung. Auf der Ebene des faktenorientierten Retrievals würde eine Anfrage nach Instanzen von High-Density-Disketten nun richtigerweise auch FLOPPY.1-01 in der Ergebnismenge ausgeben. Damit erweisen sich die auf den inkrementellen Erweiterungen von Beschreibungen zu Diskursentitäten basierenden Schlußfolgerungsprozesse als wichtiger Bestandteil des Textverstehens.

5.2.1.1 Prinzip des sprachunabhängigen ontologischen Entwurfs

Entsprechend der Anforderung, ontologisches Wissen sprachunabhängig darzustellen, orientiert sich die Modellierung in SYNDIKATE an ontologischen Kategorien und nicht an den Phänomenen des tatsächlichen Sprachgebrauchs. Die Beispielsätze 14 und 15 geben für jeweils eine Domäne eine typische Attribuierung von Eigenschaften an Objekte wieder: *Computer – Taktfrequenz* und *Patient – Gastritis*.

14. Der Computer arbeitet mit einer Taktfrequenz von 800 Mhz.

15. Der Patient hat eine chronische Gastritis.

In einem an der Sprachverwendung ausgerichteten Modellierungsansatz könnte die Wissensingenieurin/ der Wissensingenieur jedem der relevanten Konzepte, i.e. COMPUTER-

SYSTEM und PATIENT mit einer entsprechenden Beschreibung in Form einer Rolle (HAS-CLOCK-FREQUENCY bzw. HAS-DISEASE) versehen. Neben der Tatsache, daß diese Strategie zur Konzeptmodellierung ontologisch nicht adäquat ist, ergäben sich an sehr generellen Konzepten wie COMPUTER-SYSTEM eine große Anzahl von Eigenschaftsbeschreibungen, die letztlich nicht aufzählbar sind. Im Gegensatz dazu basiert die an physikalischen Gegebenheiten orientierte Modellierung auf rein ontologischen Bezügen zwischen Konzepten, die auf einer Reihe typischer Relationierungsschemata aufsetzen. Genau diese typischen Relationierungsschemata verwendet der Path-Finder (Markert, 1999), der die konzeptuellen Bezüge bei der semantischen Interpretation berechnet, um die impliziten Bezüge im Sprachgebrauch zwischen Diskursentitäten formal zu rekonstruieren (der Path Finder ist in Kap. 4.1 beschrieben).

Anhand der beiden eingeführten Beispiel wird jedoch auch offensichtlich, daß man sich an dieser Stelle ein Granularitätsproblem einhandelt. Die Konzepte COMPUTER-SYSTEM und CLOCK-FREQUENCY sind in der informationstechnologischen Wissensbasis über den Pfad HAS-CENTRAL-UNIT, HAS-MOTHERBOARD, HAS-SLOT, USES-SLOT und HAS-CLOCK-FREQUENCY relationierbar; bei PATIENT und GASTRITIS wird ein ähnlich langer Relationspfad berechnet. Abstrahierend betrachtet gilt die Feststellung, je domänenspezifischer die beiden Konzepte, umso länger der zu erwartende Relationspfad. Aus diesem Grund läßt der Path-Finder eine Parametrisierung der maximalen Pfadlänge in Abhängigkeit von der Modellierungsdichte zu.¹⁶ Auf der Ebene des Knowledge Engineering bestimmt sich die Granularität für Teilgraphen im konzeptuellen Wissen empirisch aus den analysierten Texten.¹⁷ Beispielsweise erforderte das Auftreten von Texten über Motherboards im Trainingsset die Einführung von Konzepten wie CPU-Slots oder Bussystemen. Die Verfeinerung der Domänenwissensbasis hat freilich unmittelbare Rückwirkung auf die Länge von Relationspfaden, die der Path-Finder berechnet, wodurch die Parametrisierbarkeit ihre Rechtfertigung erfährt.

¹⁶Die kontextabhängige Parametrisierung befindet sich noch im Stadium prototypischer Implementierung.

¹⁷Dies ist eine der Motivationen für die Verwendung eines Trainingssets im empirischen Teil dieser Arbeit (siehe Kap. 6.1).

5.2.1.2 Knowledge-Engineering-Prinzipien für Verbkonzepte

Der Begriff des Verbkonzepts könnte zunächst suggerieren, daß das Prinzip der sprachunabhängigen Modellierungen auf dieser Ebene keine Gültigkeit besäße. Mit diesem Begriff soll jedoch ausgedrückt werden, daß sich die konzeptuelle Repräsentation primär von einem Verb ableitet. Morphologisch von diesem Verb abgeleitete Formen wie Nominalisierungen oder deverbalisierte Adjektive greifen im allgemeinen ebenfalls auf dessen konzeptuelles Korrelat zurück (vgl. Hull & Gomez (1996)). Dadurch bleibt die Modellierung von grammatikalischen Kategorien unabhängig.

Die methodische Grundlage bei der Modellierung von Verbkonzepten entstammt den empirischen Arbeiten von Ballmer & Brennenstuhl (1986) zu den deutschen Verben. In dieser Studie, die immerhin ca. 20.000 deutsche Verben klassifiziert, erfolgt die Extrapolierung des Bedeutungskerns eines Verbs mithilfe kanonisierter Paraphrasen. Die im Analysekorpus durch Paraphrasierungen aufbereiteten Verben sind weiterhin in ca. 700 Kategorien zusammengefaßt, welche sich wiederum auf 35 Modelle verteilen. Die Kategorien gruppieren Verben mit gleichem Bedeutungskern, wohingegen die Modelle Zustands- und Aktionssequenzen beinhalten, die zumeist zeitlich oder kausal aufeinanderfolgen.¹⁸ Beispielsweise gibt es ein Buy-and-Sell-Modell, das mit dem Anbieten eines Produkts beginnt und schließlich mit dem Aushändigen des Produkts bzw. dem Empfang des Produkts endet. Zwischen diesen beiden Ereignissen steht eine (oftmals partiell geordnete) Folge von Aktionen.

Da Verben in ihrer finiten Form innerhalb eines dependentiellen Grammatikansatzes als hierarchisch dominanter Kopf einen besonderen Stellenwert besitzen, sind die an Verbkonzepten hängenden Typ-Constraints essentiell für die adäquate inhaltliche Analyse eines Einzelsatzes. Folglich sind einfachere Konzeptionen etwa in der Tradition von Schank (1975) zu grobkörnig, um tieferes Textverstehen zu ermöglichen. Die Entscheidung, das differenzierte Verbkompodium von Ballmer & Brennenstuhl (1986) als Grundlage der Verbmodellierung zu verwenden, legitimiert sich aus den konkreten Anforderungen eines tieferen Textverstehens. Jedoch ergibt sich aus deren Arbeiten nur eine ontologische Gruppierung der Verbkonzepte, ohne daß formale Basis definiert wird.

Nicht zu entnehmen sind dieser ontologischen Gruppierung die für SYNDIKATE not-

¹⁸Wegen des gemeinsamen Bedeutungskerns von Verbkategorien operieren die in Kap. 4.1.3 beschriebenen Regeln zur konzeptuellen Interpretation auf den Kategorien, wodurch sich die Regelmenge beschränken läßt.

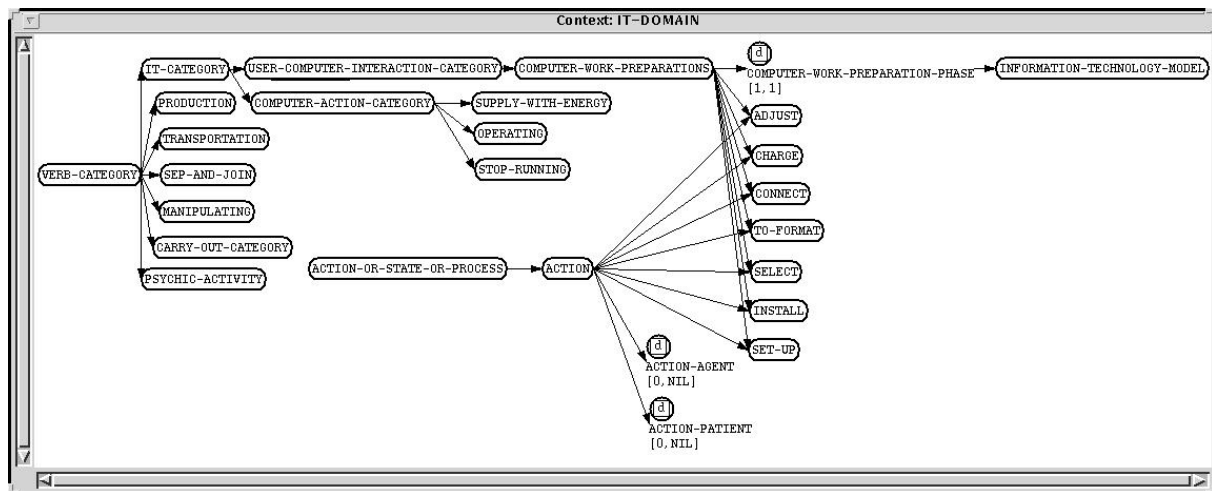


Abbildung 5.4: Fragment der Ontologie für Verbkonzepte

wendigen Beschreibungen konzeptueller Rollen, die die Verbkonzepte eigentlich erst definieren und die der eigentlichen inhaltlichen Repräsentation der syntaktischen Füller von Verbvalenzen dienen. Aus diesem Grund existiert eine weitere generische Hierarchie im Teilgraphen der Verbkonzepte. Grundsätzlich wird differenziert, ob ein Verbkonzept eine Handlung (ACTION), einen Zustand (STATE) oder einen Prozeß (PROCESS) bezeichnet. Eine Handlung zum Beispiel definiert sich dadurch, daß es einen Agens (Ausführenden) und einen Patiens (betroffenes Objekt) gibt. Je nach Situation können weitere konzeptuelle Beschreibungen hinzugefügt werden. Das Verbkonzept für „geben“ hält einen Slot für den Rezipienten (Empfänger eines Objekts) vor. Allen Verbkonzepten gemein ist, daß sie jeweils mit Zeit- und Ortsangaben in Verbindung gebracht werden können.

Die in die 80-er Jahre zurückdatierenden empirischen Arbeiten von Ballmer & Brennenstuhl (1986) lassen aufgrund des technologischen Wandels Analysen zu informationstechnologischen Verben vermissen. Dennoch bieten die vorgegeben Modelle und Kategorien Anhaltspunkte für das Ontological Engineering. Das in Abb. 5.4 enthaltene Fragment der Ontologie für informationstechnologische Verbkonzepte veranschaulicht, wie die drei Beschreibungsebenen für Verbkonzepte, VERB-MODEL, VERB-CATEGORY und ACTION-OR-STATE-OR-PROCESS, zusammenhängen.

Im linken Teil der Abbildung befindet sich das Konzept VERB-CATEGORY mit einer Reihe von Unterkonzepten, von denen eines, nämlich IT-CATEGORY, detaillierter zur Darstellung kommt. Das Unterkonzept COMPUTER-WORK-PREPARATIONS subsumiert die-

jenigen Verbkonzepte, welche die Handlungen modellieren, die ein Computer-Benutzer zur Vorbereitung der Interaktion mit dem Rechner vornimmt. Das Konzept ist außerdem über die Relation `COMPUTER-WORK-PREPARATION-PHASE` als eine Phase innerhalb des `INFORMATION-TECHNOLOGY-MODELS` definiert (Abb. 5.4 rechts oben). Neben dieser kategorialen Systematik existiert die an Kasusrollen orientierte, die in der Abbildung durch das Konzept `ACTION` eingeführt wird. Alle Verbkonzepte, die Unterkonzepte der Kategorie `COMPUTER-WORK-PREPARATIONS` sind, werden ebenfalls von `ACTION` subsumiert. Von diesem Oberkonzept erben sie die Rollen `ACTION-AGENT` (für das syntaktische Subjekt) und `ACTION-PATIENT` (für das syntaktische Objekt). Weiterhin verdeutlicht dieses Fragment der Ontologie die Notwendigkeit multipler Vererbung in der Wissensbasis, ohne die unterschiedliche Modellierungssichten auf Konzepte nicht realisierbar wären.

Zum gegenwärtigen Zeitpunkt finden sich in `SYNDIKATE` insgesamt 39 Verbmodelle, ca. 60 Verbkategorien und ca. 200 Verbkonzepte. Darüber hinaus gibt es 25 auf Kategorien basierende konzeptuelle Interpretationsregeln (insgesamt etwas über 50). Zu beachten ist, daß ein Verbkonzept zugleich auf mehrere Verben bzw. auf von Verben abgeleitete Wortformen abgebildet werden kann. Folglich decken die Verbkonzepte im Lexemlexikon ein Mehrfaches an Verbeinträgen ab.

Mit den erwähnten konzeptuellen Relationen `AGENT` und `PATIENT` ist der auf der Ebene der Verbkonzepte wichtige Aspekt des Inventars und der Organisation thematischer Rollen eingeführt. Bei der Verbmodellierung finden konzeptuelle Relationen Verwendung, die sich an den ontologischen Konzeptionen von Nirenburg & Raskin (1987) und Quantz et al. (1994) orientieren. Je unterschiedliche Teilmengen von diesen sind über die statische semantische Interpretation mit Dependenzrelationen verbunden (vgl. 4.1.1). Neben den allgemeinen Relationen, wie etwa `AGENT`, `PATIENT` oder `BENEFICIENT` existieren Spezialisierungen, um die an spezielle Verbkonzepte geknüpften Restriktionen repräsentieren zu können. Beispielsweise verfügt das Konzept `ACTION` über die Rolle `ACTION-PATIENT`, die von `PATIENT` subsumiert wird. Schließlich besitzt jedes Unterkonzept zu `ACTION` eine eindeutige Relation für den thematischen `PATIENT`. Das Konzept `INSTALL` etwa definiert für die Rolle `INSTALL-PATIENT` einen Wertebereich, der auf `HARDWARE` oder `SOFTWARE` beschränkt ist. Auf diese Weise stellt die Wissensbasis die zur inhaltlichen Analyse benötigten feinkörnigen Integritätsbedingungen bereit.

5.2.1.3 Partonomien in der medizinischen Wissensbasis

Eine domänenspezifische Besonderheit stellt die Bedeutung partitiver Hierarchien innerhalb der Medizin dar. Zum einen ergibt sich die Relevanz der PART-OF-Beziehungen aus der Wichtigkeit lokaler Bezüge innerhalb der für die Medizin fundamentalen Anatomie. Zum anderen setzen medizinisch bedeutsame Schlußfolgerungsmuster auf Partonomien auf: die Koordination von Taxonomien und die Propagierung von Rolleneigenschaften entlang von PART-OF-Hierarchien (Rector & Horrocks, 1997).

Gegeben seien zwei konzeptuelle Relationen REL_1 and REL_2 , von denen REL_2 als Subrelation zu PART-OF definiert ist. Eine Rolleneigenschaft wird dann propagiert, wenn die nachfolgende Implikation erfüllt ist, wobei CON_1 bis CON_3 Konzeptbezeichner seien:

$$CON_1 \text{ REL}_1 CON_2 \wedge CON_2 \text{ REL}_2 CON_3 \Rightarrow CON_1 \text{ REL}_1 CON_3$$

Wenn ein weiteres Konzept CON_4 existiert, das mit CON_3 über REL_2 relationiert ist, dann gilt auch $CON_1 \text{ REL}_1 CON_4$, weil $CON_1 \text{ REL}_1 CON_3$ bereits abgeleitet wurde. D.h., dieses Schlußmuster gilt für jedes Konzept CON_n , welches transitiv über PART-OF mit CON_1 relationiert ist. Aus diesem logischen Sachverhalt resultiert der Name Rollenpropagation, da REL_1 gewissermaßen in Gegenrichtung zur partitiven Hierarchie “vererbt” wird.

Die Koordination von Taxonomien entlang von PART-OF Relationen stellt eine Spezialform der Rollenpropagation dar, indem der Antezedent der Implikation um eine Relationierung erweitert wird. Es existiere noch ein Konzept CON_4 , welches mit REL_1 zu CON_3 relationiert sei. Als Konsequenz läßt sich ableiten, daß CON_4 das Konzept CON_1 subsumiert:

$$CON_1 \text{ REL}_1 CON_2 \wedge CON_4 \text{ REL}_1 CON_3 \wedge CON_2 \text{ REL}_2 CON_3 \Rightarrow CON_1 \text{ Is-A } CON_4$$

Zur Veranschaulichung der beiden Schlußfolgerungsmuster enthält die Abb. 5.5 ein Beispiel in Anlehnung an Rector et al. (1997).¹⁹ Das Konzept “Fraktur des Oberschenkelschafts” FRACTURE-OF-SHAFT-OF-FEMUR ist mit dem anatomischen Teil “Oberschenkelschaft” SHAFT-OF-FEMUR über FRACTURE-OF relationiert (vgl. Abb. 5.5 ❶). Zugleich hat der Oberschenkelknochen (FEMUR) das anatomische Teil Oberschenkelschaft (SHAFT-

¹⁹Die Grafik ist entnommen aus Hahn et al. (1999c).

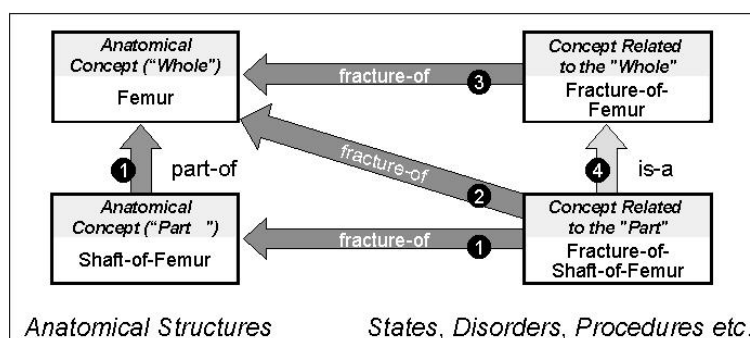


Abbildung 5.5: Reasoning auf partitiven Relationen

OF-FEMUR), ausgedrückt durch die PART-OF Relation zwischen beiden Konzepten (Abb. 5.5 ❶). Aus der Regel für Rollenpropagation folgt, daß eine Fraktur des Oberschenkelknochens auch eine Fraktur des Oberschenkelknochens darstellt (Abb. 5.5 ❷). Das durch die Einser-Ziffern gegebene Pattern wird nun dahingehend erweitert, daß sich eine Fraktur des Oberschenkelknochens FRACTURE-OF-FEMUR durch die Relation FRACTURE-OF mit dem Wertebereich FEMUR definiert (Abb. 5.5 ❸). Gegeben die partitive Relation zwischen FEMUR und SHAFT-OF-FEMUR läßt sich nun die Subsumptionsbeziehung der Konzepte FRACTURE-OF-FEMUR und FRACTURE-OF-SHAFT-OF-FEMUR inferieren (Abb. 5.5 ❹).

Obwohl unbestritten ist, daß die auf der Transitivität der PART-OF-Relation aufbauenden Partonomien eine herausragende Rolle in der medizinischen Domäne spielen, sind die dem Reasoning unterliegenden Prinzipien noch nicht ausreichend geklärt (vgl. auch den Überblicksartikel von Artale et al. (1996)). Insbesondere die von Winston et al. (1987) vertretene Hypothese, daß die allgemeine PART-OF-Relation nicht transitiv sei, wohl aber deren Subrelationen, solange keine zweite PART-OF-Beziehung in einer Relationierungskette anzutreffen ist, kann in der Medizin empirisch widerlegt werden (Bernauer, 1996). Jedoch steht eine umfassende empirische Studie zur Kombinierbarkeit von unterschiedlichen PART-OF Relationen für die Medizin noch aus (vgl. hierzu die domänenunabhängigen Arbeiten von Sattler (1995)).

Darüber hinaus scheinen die Schlußfolgerungsmuster der Rollenpropagation wie auch der taxonomischen Koordination nicht grundsätzlich zu gelten. In Abb. 5.6 findet sich ein Beispiel hierzu. Die Appendix ist unbestritten ein Teil (wenn auch nur ein "Anhang") des Dickdarms. Eine Perforation der Appendix stellt –konform dem Schlußmuster der Rollenpropagation– eine Perforation des Dünndarms dar. Jedoch kann von einer Appen-

ditis (Entzündung des Appendix) logisch nicht auf eine Kolitis (Entzündung des Darms) geschlossen werden. Es wäre möglich, daß die Rollenpropagation eine feste Eigenschaft von konzeptuellen Relationen (im vorliegenden Beispiel also von INFLAMMATION-OF, nicht aber von PERFORATION-OF) ist; in diesem Sinne gebrauchen Rector & Horrocks (1997) den in ihre terminologische Sprache eingeführten *specializes*-Operator, der die Eigenschaft der Rollenpropagation fest mit der Relation verbindet. Wie Schulz et al. (1998) jedoch gezeigt haben, erlaubt ein und dieselbe konzeptuelle Relation (z.B. INFLAMMATION-OF) in einigen Fällen die Part-Whole-Spezialisierung, in anderen Fällen hingegen nicht.

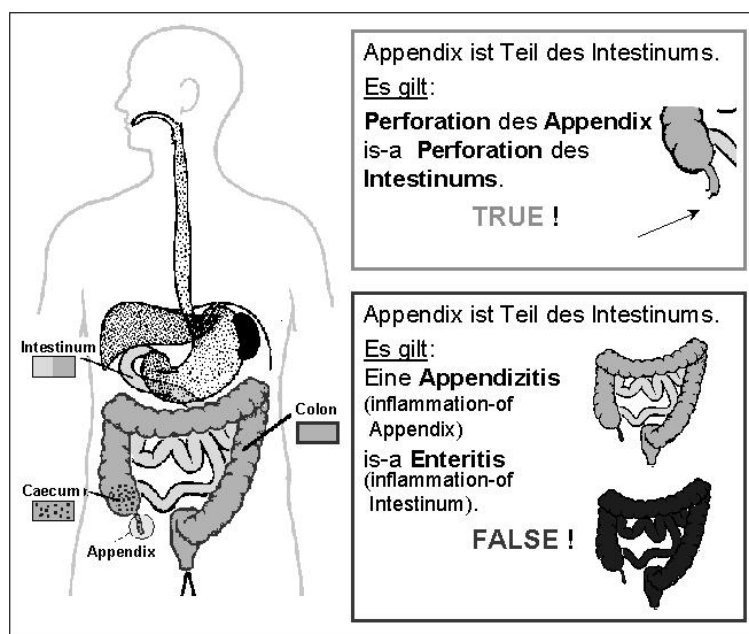


Abbildung 5.6: Eingeschränkte Generalisierbarkeit partitiver Schlußfolgerungsmuster

Um die in diesem Kapitel beschriebenen und auf partitiven Hierarchien basierenden Reasoningmuster zu ermöglichen sowie um die damit verbundenen Modellierungsprobleme zu lösen, ohne transitive Relationen in die terminologische Sprache einzuführen, verwendet die medizinische Wissensbasis in MEDSYNDIKATE ein besonderes formales Konstrukt, die sog. SEP-Triplets (Schulz et al., 1998). Diese bestehen aus einem P-Knoten, der sämtliche Teile eines Konzepts subsumiert und einem E-Knoten für die gesamte Entität. Beide Konzepte werden wiederum von einem besonderen S-Knoten für die gesamte Struktur subsumiert. Durch die Einführung zusätzlicher partitiver Konzepte in die generische Hierarchie emuliert die Subsumptionsbeziehung die Transitivität der PART-OF-Relation (vgl. hierzu

Hahn et al. (1999b)). Im Gegensatz zu Schmolze & Mark (1991), die nur einen zusätzlichen Knoten pro Konzept einführen, erweitert das SEP-Schema eine Konzeptbeschreibung um zwei Knoten. Dieses dritte Konzept, der S-Knoten, dient insbesondere dazu, das flexible Ein- und Ausschalten der Koordination von generischen Hierarchien auf der Grundlage von PART-OF-Beziehungen zu ermöglichen (Hahn et al., 1999c). Darüber hinaus erlaubt die Unterscheidung zwischen einem Entitätsknoten, der das Ganze repräsentiert, und dem P-Knoten, als dem Oberkonzept aller Teile, eine ontologisch sinnvolle Differenzierung beim Ontological Engineering.

Die Bedeutung der domänenadäquaten Modellierung innerhalb der Medizin ergibt sich notwendig aus dem für SYNDIKATE angestrebten Faktenretrieval. Im Fall einer Anfrage zu Brüchen des Oberschenkelknochens müssen in der Menge der Antworten auch sämtliche Instanzen von Frakturen des Oberschenkelschafts enthalten sein. Ohne Koordination von taxonomischen Hierarchien entlang von meronymischen Beziehungen wäre dies nicht möglich. Abschließend sei betont, daß die herausgearbeiteten formalen Prinzipien auf andere Domänen übertragbar sind, wenn die ontologische Struktur des betrachteten Welt-ausschnitts dies erfordert.

5.3 Wissensrepositorien

Da Wissen in allen Bereichen intelligenter Verfahren zur Textindexierung, zur Textklassifikation wie auch zum Textverstehen eine zunehmend wichtigere Rolle spielt, gibt es sowohl allgemeingültig enzyklopädische als auch dediziert domänenabhängige Wissensrepositorien.²⁰

Die Entwicklung und Wartung großer Ontologien ist eine zeitaufwendiges und teures Unternehmen. Andererseits macht die Abstützung von Applikationen durch ontologisches Wissen nur dann Sinn, wenn dieses in ausreichendem Umfang zur Verfügung steht. Daraus folgt, daß die Erstellung und die Pflege großer Wissensrepositorien nur von solchen Institutionen bzw. Firmen unternommen werden kann, die über ausreichende personelle Ausstattung über einen längeren Zeitraum verfügen. Als Beispiele wären im kommerziellen Sektor die CYC-Corporation zu nennen, im öffentlichen die *National Library of Medicine*.

²⁰Die Verwendung des Begriffs ‘Wissensrepositorium’ ist bewußt neutral und umfaßt ein weites Spektrum von einfachen Thesauri über feinkörnige, semantisch reichhaltige Ontologien bis hin zu formal fundierten, tief modellierten Konzeptsystemen.

Da jedoch diese Ressourcen – sofern überhaupt frei verfügbar – nicht auf die Anforderungen eines bestimmten Systems zugeschnitten sind, stellt sich die Frage, ob und wie das in ihnen enthaltene Wissen für die Zwecke von SYNDIKATE genutzt werden kann.

Alternativ dazu gibt es Bestrebungen, über definierte Schnittstellensprachen zwischen Ontologien Wissen auszutauschen. Mit dem KIF (Knowledge Interchange Format) gibt es eine Sprache, über die basierend auf einer Logik erster Ordnung und nicht-monotonen Schlußfolgerungsregeln Wissen aus Wissensbasen importiert und exportiert werden kann (Neches et al., 1991). Ontolingua bietet darüber hinaus noch eine frame-artige Wissensrepräsentationssprache, um Ontologien zu definieren (Gruber, 1993). Da jedoch in der Praxis keines der im folgenden vorgestellten Wissensrepositorien einen dieser Standards benutzt, kann diese – zumindest von der theoretischen Seite aus betrachtet – überzeugende Konzeption leider nicht zum Einsatz kommen, um die SYNDIKATE-Wissensbasen anzureichern.

Einen kurzen Überblick über domänenunabhängige Wissensrepositorien gibt das nachfolgende Kapitel unter dem Gesichtspunkt, inwieweit das in ihnen kodierte Wissen die Entwicklung der SYNDIKATE-Wissensbasen zu unterstützen in der Lage ist. Anschließend folgt eine Übersicht über die wesentlichen Wissensrepositorien in der Medizin.

5.3.1 Domänenunabhängige Wissensrepositorien

WORDNET ist ein lexikalisches Referenzsystem, das sich an psycholinguistischen Theorien orientiert und Evidenzen psycholinguistischer, und dabei teilweise sogar sprachpathologischer Untersuchungen berücksichtigt (Miller et al., 1990). Die Kernstruktur des lexikalischen Referenzsystems ist durch die Wortkategorien Nomen (Miller, 1998), Verben (Fellbaum, 1998a) und Adjektive (Gross & Miller, 1990) vorgegeben, damit werden folglich keine sprachunabhängigen ontologischen Kategorien gebildet. Sogenannte *synsets* (synonym sets) dienen der Repräsentation bedeutungsgleicher Wörter. Mehrdeutigkeiten eines Worts werden entsprechend durch die Präsenz in mehr als einem *synset* dargestellt. In der gegenwärtig aktuellen Version 1.5 verfügt WORDNET über 91.600 *synsets*, von denen gut zwei Drittel auf Nomen entfallen (Fellbaum, 1998b).

Die Art und Weise, wie SYNDIKATE im Lexikon Wortoberflächen und Konzeptidentifizier verwaltet, läßt sich ohne weiteres mit der Organisation von *synsets* in WORDNET vergleichen. Im Fall von Polysemien reflektiert die zu einem Lexem definierte Menge an Konzeptidentifiern die Anzahl der unterschiedlichen Lesarten. Umgekehrt läßt sich die

Menge von synonymen Wörtern zu einer Wortbedeutung (repräsentiert durch einen eindeutigen Konzeptidentifizier) dadurch bestimmen, daß alle lexikalischen Einträge mit dieser Wortbedeutung zusammengefaßt werden. Eine Unterscheidung von Lexem und Konzept wie in SYNDIKATE gibt es jedoch nicht.

Obwohl die Referenzen zwischen lexikalischen Einheiten nur auf schwachen semantischen Relationen basieren (im Großen und Ganzen Synonymie, Antonymie, Hyponymie und Meronymie, vgl. Miller (1995)), lassen sich wichtige Modellierungshinweise für das Knowledge Engineering aus WORDNET gewinnen. Zum einen sind die an Wortbedeutungen hängenden semantischen Wohlgeformtheitsbedingungen bezüglich ihrer Verwendbarkeit in sprachlichen Kontexten in terminologische Axiome übertragbar. Zum anderen liefert die Kategorienbildung im oberen Bereich der Hypernymie- bzw. Hyponymie-Beziehungen wichtige Orientierungspunkte für das UPPER-MODEL in SYNDIKATE.

Die PENMAN Wissensbasis (Bateman et al., 1990) markiert in der Genealogie von Ontologien die Vorgängerin einer Reihe von Weiterentwicklungen. Die wichtigsten Abkömmlinge sind die terminologische Taxonomie SENSUS²¹ und das Generalized Upper Model (GUM) in der Version 2 (Bateman et al., 1995). Während SENSUS als Weiterentwicklung von PANGLOSS (Knight & Luk, 1994) Wissen aus WORDNET wie auch aus CYC integriert, ging das GUM aus der Mischung mehrerer Ontologien hervor (Henschel & Bateman, 1994). Auf der Ebene der Abstraktion von den linguistischen Realisierungen an der sprachlichen Oberfläche hin zu sprachunabhängigen ontologischen Konzeptrepräsentationen nimmt das GUM – ähnlich dem UPPER-MODEL in SYNDIKATE – eine Mittelstellung ein (Bateman et al., 1995). Als Wissensrepräsentationssprache dient dem GUM eine um undokumentierte Makros erweiterte Variante von LOOM (MacGregor, 1994), beispielsweise zur Reifikation von Relationsbezeichnern (MacGregor, 1993), womit insbesondere Constraints für den Domänen- und Rangebereich von Relationsfüllern definiert sind. Gleichwohl wird in den Konzept- und Relationsdefinitionen bei weitem nicht die Ausdrucksmächtigkeit von LOOM ausgeschöpft.

Das UPPER-MODEL von SYNDIKATE orientiert sich bei den Top-Level Konzepten am GUM mit Ausnahme der Modellierung von Verbkonzepten, die – wie bereits oben beschreiben (vgl. Kapitel 5.2.1.2) – auf einer mehrdimensionalen Sicht im Sinne einer multiplen Vererbungshierarchie beruhen. Die breite Differenzierung in der Relationenhierarchie, die

²¹Zu SENSUS gibt es leider keine gedruckte Publikation, statt dessen sei hier auf die Web-Seite verwiesen: <http://www.isi.edu/natural-language/projects/ONTOLOGIES.html>.

in SYNDIKATE bedingt durch die domänenspezifischen Erweiterungen noch tiefer reicht, in Verbindung mit an den Relationsdefinitionen kodierten Constraints finden bei der Modellierung des UPPER-MODELS ebenfalls Berücksichtigung. Interessanterweise enthält das Generalized Upper Model auch eine Menge an logischen Relationen, die im Zuge der RST (Mann & Thompson, 1988) zur Annotation rhetorischer Strukturen in Texten Verwendung finden.

Inzwischen kommerzialisiert hat sich der letzte Vertreter, der in diesem Kapitel vorgestellten domänenunabhängigen Wissensrepositorien, der CYC Knowledge Server (Lenat & Guha, 1990; Guha & Lenat, 1990). Das in CYC kodierte Wissen umfaßt Fakten, die in verschiedene Mikrotheorien zusammengefaßt sind, mit mittlerweile weit über eine Million Einträgen sowie Heuristiken für das Schlußfolgern über Objekte und Events (Lenat, 1995). Formal ist das Wissen in einer eigenen, an Prädikatenlogik erster Ordnung ausgerichteten frame-artigen Wissensrepräsentationssprache, CYCL (Lenat & Guha, 1990), gefaßt, die um eine Constraint-Sprache und automatische Deduktionsmechanismen wie Vererbung und Klassifikation erweitert ist. Um bei der immensen begrifflichen Abdeckung dennoch über einen sauberen ontologischen Entwurf zu verfügen, wurde bei der Entwicklung von Anfang an auf eine klare Kategorienbildung auf der obersten Ebene der Konzepthierarchie geachtet (Barnett et al., 1990). Im Gegensatz zu den Spezialontologien ist die Top-Level-Ontologie trotz des kommerziellen Charakters von CYC öffentlich zugänglich. Aufgrund der langjährigen Erfahrung der CYC-Gruppe im Ontological Engineering orientiert sich das UPPER-MODEL von SYNDIKATE an CYC, insbesondere in den Teilen, die nicht als Oberkonzepte der Spezialontologien MED und IT dienen.

Den allgemeinen Modellierungstiefe im Upper-Model und von enzyklopädischen Konzepten zum trotz warten alle domänenunabhängigen Wissensrepositorien – sofern frei zugänglich – mit vergleichsweise wenig Konzepten zu den in SYNDIKATE verarbeiteten Spezialdomänen auf. Darüber hinaus reicht die Granularität der Modellierung bei weitem nicht aus, um ein wirklich tieferes Textverstehen zu ermöglichen. Während für die Informationstechnologie keine dedizierten Ontologien zur Verfügung stehen, finden sich in der Medizin eine Vielzahl von Wissensrepositorien, die sich bei der domänenspezifischen Modellierung als nützliche Grundlage anbieten; und dies sowohl bezüglich der Abdeckung als auch bezüglich der Modellierungstiefe.

5.3.2 Wissensrepositorien in der Medizin

Neben der Chemie und der mit ihr verbundenen Pharmakologie ist die Medizin die Fachsprache mit den meisten Termini. Das fachspezifische Vokabular aller medizinischen Disziplinen dürfte in der Größenordnung von weit über einer Million Terme liegen (cf. Humphreys et al. (1998)). Sich aus der Organisation des Gesundheitswesens heraus ergebende Zwänge zu statistischen Erhebungen, klinischer Dokumentation und fallbasierter Abrechnung bei den Kostenträgern führen zu einer breiten Diskussion über die Normierung des medizinischen Vokabulars (Friedman et al., 1995b; Cimino, 1998b; 1998c). Die heute öffentlich verfügbaren Ressourcen medizinischer Terminologie sind Ausdruck und Ergebnis der Bemühungen in dieser Richtung. Darüber hinaus existieren – zumindest innerhalb Europas – Teilbereiche medizinischen Wissens, die einer konkreten Normierung unterliegen (CEN/TC251/WG2/PT003, 1995). Die Wissensrepositorien, welche in diesem Unterkapitel kurz beschrieben werden, sind *Thesauri* für die Indexierung medizinischer Texte (MESH), *Kodierungssysteme* für Diagnosen, Prozeduren und Befunde (ICD-9, ICD-10, ICPM und READCODES), kontrollierte *Vokabularien* für die klinische Dokumentation (SNOMED-3) sowie das anatomische Modell im University of Washington Digital Anatomist (UWDA). Eine Sonderstellung nimmt das UMLS ein (McCray & Razi, 1995), da es als Metathesaurus mittlerweile über 50 unterschiedliche Wissensrepositorien zusammenfaßt einschließlich der im Vortext aufgezählten Referenzen.²²

Einen Thesaurus mit multiplen Begriffshierarchien zur Indexierung von Zeitschriftenartikeln stellt MESH (Medical Subject Heading) dar (NLM, 1997). Insbesondere die in der Medizin für das Retrieval von Fachartikeln wichtige bibliographische Datenbank MEDLINE verwendet Indexterme aus MESH. Anstelle von generischen Begriffshierarchien finden sich zwischen Termen nur die schwachen semantischen Beziehungen BROADER-TERM und NARROWER-TERM, was im Resultat zu einer Vermischung partitiver und generischer Hierarchien führt. Zum Beispiel “subsumiert” BLOOD sowohl PLASMA, ein Blutbestandteil und damit in einer meronymischen Beziehung zu PLASMA stehend, und FETAL-BLOOD, welches BLOOD tatsächlich spezialisiert.

Das zum aktuellen Zeitpunkt maßgebliche System zur Verschlüsselung von Diagnosen ist die ICD-10 (*International Classification of Diseases* (ICD-10, 1992)), welches die ältere ICD-9 mit einer Vielzahl von Veränderungen abgelöst hat (Zaiß et al., 1996). ICD-

²²Ein Überblick zu den Wissensrepositorien findet sich in (Cimino, 1995).

10 besteht aus einem mono-axialen, hierarchischen Beschreibungssystem zur Vergabe von Diagnoseschlüsseln. Im Gegensatz zu SNOMED, dessen Achsen frei kombinierbar sind, gibt ICD-10 Schlüsselkombinationen fest vor. Neben der Verknüpfung von Beschreibungsdimensionen verwendet ICD-10 generische Hierarchien zur Strukturierung seines Wissensbestands. ICD-9-CM stellt eine dem US-amerikanischen Gesundheitswesen angepaßte Erweiterung dar, die zudem um eine Prozeduren-Klassifikation ergänzt ist. Eine besondere Klassifikation existiert darüber hinaus für Tumorerkrankungen (ICD-0). Auch in der ICD-10 finden sich die den Beschreibungssystemen ohne formale Fundierung inhärenten Probleme etwa in Form inkonsistenter Konzeptbeschreibungen.

Eine Zwischenstellung in den medizinischen Wissensrepositorien nehmen die READ CODES ein (O'Neil et al., 1995). Einerseits als Thesaurus und Kodierungssystem konzipiert, verfügen die READ CODES über eine multiple hierarchische Struktur, die sowohl generische als auch partitive Beziehungen zwischen Konzepten modelliert. Durch die Erweiterungen um ein *template*-Schema, mit dem sich kompositional Attribute zu Konzepten beschreiben lassen, bilden die Konzeptbeschreibungen ein semantisches Netz (Schulz et al., 1997). Insgesamt sind über 7.000 anatomische Konzepte im semantischen Netz mit 16.000 operativen Prozeduren und 40.000 pathologischen Störungen relationiert. Aufgrund ihrer in Ansätzen kompositionalen Beschreibungssprache finden READ CODES weitergehende Verwendung für beispielsweise die semantische Definition von Krankheiten (Brown et al., 1998).

Die SNOMED (Systemized Nomenclature in Medicine (Coté, 1993)) dient als kontrolliertes Vokabular der Standardisierung in der medizinischen Dokumentation. Die Kombination der verschiedenen Achsen in SNOMED ermöglicht die Konstruktion kompositionaler Konzepte, wodurch SNOMED in einer empirischen Untersuchung zu medizinischen Wissensrepositorien den höchsten Grad an Vollständigkeit zugesprochen bekam (Campbell et al., 1997). Leider ist es möglich, mit SNOMED, bedingt durch die semantische Unspezifikation der Beschreibungssprache, unsinnige Konzeptdefinitionen aufzubauen (z.B. Fraktur des Bluts). Darüber hinaus erlaubt SNOMED redundante Kodierungen desselben Konzepts durch verschiedene Codes und vermischt ebenfalls generische und partitive Hierarchien (Ingnerf & Diedrich, 1997).

Das im UWDA modellierte anatomische Modell (Rosse et al., 1998b) verwendet eine reiche Beschreibungssprache, um anatomische Konzepte, aber auch konzeptuelle Beziehungen zwischen diesen zu modellieren. Die Anzahl der anatomischen Konzepte, die auf der Grundlage eines anatomischen Strukturmodells implementiert sind, liegt mittlerweile

bei deutlich über 25.000 (Mejino & Rosse, 1999) mit einem Schwerpunkt auf die Thorax-Domäne. Einen besonderen Schwerpunkt in der ontologischen Konzeption der Anatomie bilden spatiale Konzepte sowie deren räumliche Beziehungen zueinander. Die konzeptuelle Relationierung der spatialen Entitäten erfolgt entlang von drei Dimensionen: der Topologie, der partonymischen Beziehungen sowie einem räumlichen Assoziationsnetzwerk (Neal et al., 1998). Zwar liegen die Konzeptdefinitionen im Rahmen eines formal schwachen semantischen Netzwerks vor, jedoch richten die Entwickler der UWDA ihr besonderes Augenmerk auf die ontologische Konsistenz der anatomischen Domäne. Beispielsweise benutzt Mejino & Rosse (1998) den UWDA zur Konsistenzprüfung im UMLS.²³

In diesem Zusammenhang gilt es zu unterscheiden zwischen der ontologischen Konsistenz auf der einen Seite und der formalen Konsistenz auf der anderen Seite. Eine ontologische Inkonsistenz liegt beispielsweise dann vor, wenn Kategorien wie Taxonomien und Partonomien bei der Wissensmodellierung vermischt werden. Dennoch können derartige Modelle formal konsistent sein. Formale Inkonsistenzen ergeben sich im Falle von Konzept- oder Relationsbeschreibungen, die im Rahmen eines logikbasierten Paradigmas bezüglich ihrer Semantik unerfüllbar wären. Im Unterschied zu Wissensrepräsentationssystemen, die eine deduktive Reasoningkomponente verwenden, bleibt die Prüfung auf ontologische Konsistenz eine intellektuelle Leistung der Wissensingenieure.

Das Unified Medical Language System (UMLS) inkorporiert als Metathesaurus eine ständig wachsende Anzahl von in der Medizin verfügbaren Wissensrepositorien (McCray & Razi, 1995). Unter diesen befinden sich auch die meisten der in diesem Unterkapitel beschriebenen Wissensrepositorien: MESH, SNOMED, UWDA, ICD und READ CODES. Durch die Integration hinsichtlich begrifflicher Abdeckung und deskriptiver Präzision sehr unterschiedlicher Wissensquellen propagieren sich die in diesen Wissensquellen enthaltenen Unzulänglichkeiten in den UMLS-Metathesaurus (Bodenreider et al., 1998a). Neben dem reinen Zusammenfassen von Wissensrepositorien und der Bereitstellung umfangreicher Tools zu deren Nutzung nimmt UMLS zusätzlich eine semantische Typung mithilfe von Konzepten und Relationen aus einem semantischen Netzwerk vor.²⁴ Insgesamt umfaßt die Ausgabe des UMLS von 1998 476.313 Konzepte, wodurch in etwa 1,1 Millionen Terme

²³In diesen Bereich gehört ebenfalls die Arbeit von Cimino (1998a), der das UMLS ebenfalls mit ontologisch-semantischen Methoden überprüft.

²⁴Das semantische Netz stellt insgesamt 133 Konzepte und 50 Relationen zur Verfügung (McCray & Nelson, 1995).

beschrieben sind (Hunaying et al., 2000). Die semantische Typung dieser Konzepte bildet in Verbindung mit den semantischen Relationierungen, wie sie im UMLS beinhaltete Wissensrepositorien bereitstellen (z.B. durch den UWDA), die Grundlage für den Export von Konzeptbeschreibungen in formal fundierte Wissensrepräsentationssysteme (cf. Schulz & Hahn (2000)). Hinsichtlich der terminologischen Abdeckung nähert sich der UMLS asymptotisch einer weitgehenden Vollständigkeit, was durch eine größere Studie empirisch nachgewiesen werden konnte (Humphreys et al., 1998). Jedoch beeinträchtigen die semantischen Inkonsistenzen die Verwendbarkeit von UMLS in konkreten Applikationen (Bodenreider et al., 1998b).

Einen Versuch, die inhaltlichen Defizite in den großen Wissensrepositorien anzugehen, stellt die Entwicklung des MED (Medical Entity Dictionary) dar (Cimino et al., 1994). Gemäß der programmatischen Forderung, die vorkommenden Redundanzen, Mehrdeutigkeiten und auch Inkonsistenzen durch eine stärkere formale Fundierung zu beseitigen, formt das MED ein semantisches Netz zur konsistenten Repräsentation medizinischer Terminologie mit fast 33.000 Konzepten. Obwohl die Reasoning-Komponente des Systems in der Lage ist, aufgrund von Konzeptbeschreibungen Subsumptionbeziehungen zu berechnen, fehlt dem Ansatz eine fundierte formale Semantik.

Gerade der Anspruch, die Abbildung medizinischen Wissens auf eine solide formale Grundlage zu stellen (Campbell et al., 1994), führt zur zunehmenden Verwendung logikbasierter Wissensrepräsentationssysteme. Dabei kommen die Wissensrepräsentationssysteme auf zwei Ebenen ins Spiel. Erstens dienen sie der rigiden Überprüfung und Verbesserung bestehender Wissensrepositorien, indem die in den Wissensrepositorien enthaltenen Konzeptbeschreibungen auf ihre formale Konsistenz untersucht werden. Dazu sind die informellen Konzeptdefinitionen zunächst in die Syntax der Zielwissensrepräsentationssprache zu übertragen. Anschließend erfolgt die Überprüfung dieser Definitionen innerhalb des entsprechenden Wissensrepräsentationssystems. Dadurch aufgedeckte formale Modellierungsfehler können dann korrigiert werden. Auf diese Art und Weise testeten beispielsweise Rogers et al. (1998) die in den READ CODES enthaltenen Definitionen mit dem Wissensrepräsentationssystem GRAIL. Zweitens finden sie von vorneherein als Ausgangspunkt beim Knowledge Engineering Anwendung, d.h., die zu modellierende Wissensbasis wird sukzessive im Rahmen eines Wissensrepräsentationssystems erstellt (z.B. in GALEN (Rector et al., 1995b)).

Die wie K-REP (Mays et al., 1996) auf terminologischer Logik basierende Wissens-

repräsentationssprache GRAIL orientiert sich in ihrem Sprachumfang besonders an den empirisch begründbaren Notwendigkeiten des Knowledge Engineering in der Medizin (Rector et al., 1997). Daher verzichtet man insbesondere auf die Negation und die Disjunktion sowie Kardinalitätsrestriktionen für Rollen in der Konzeptdefinitionssprache, um statt dessen auf der TBox die Transitivität von Rollen und das damit verbundene Reasoning zu implementieren. Insbesondere unterstützt der Classifier die Berechnung von Subsumptionsbeziehungen entlang von transitiven Relationen, syntaktisch durch den bereits erwähnten besonderen *specializes*-Operator ausgedrückt (vgl. Kap. 5.2.1.3). Darüber hinaus hat man auf die Implementierung einer ABox verzichtet. Referenzen auf Individuen und Relationierungen zwischen diesen finden sich extern in Datenbanken (Rector & Horrocks, 1997). Damit erfolgt kein Reasoning über sich verändernde Konzeptzugehörigkeiten von Instanzen aufgrund von kompositionalen Erweiterungen der Instanzbeschreibungen – ein für das Textverstehen in SYNDIKATE essentielles Reasoningmuster.

Die im Zuge des europäischen GALEN-Projekts (Rector & Nowlan, 1993) mit GRAIL realisierte Domänenwissensbasis verfügt mittlerweile über ca. 10.000 Relationen und Konzepte. Als sprachunabhängiger konzeptueller Kern eines terminologischen Servers dient diese Wissensbasis einer Reihe von Anwendungen wie etwa der strukturierten Dateneingabe zu Patientenbeobachtung (Rector et al., 1995a).

In der Medizin zeichnet sich in den letzten Jahren ein interessanter Trend ab. Die verfügbaren Wissensrepositorien lassen sich grob geschieden in zwei Klassen kategorisieren. Auf der einen Seite existieren Vertreter wie MESH oder SNOMED, die eine große Anzahl an Termen bzw. Konzepten abdecken. Sie verfügen jedoch nur über ein sehr eingeschränktes Vokabular an semantischen Primitiven zur Beschreibung der Domäne. Darüber hinaus besitzt die den Term- bzw. Konzeptbeschreibungen zugrundeliegende Sprache keine formale Semantik, womit in diesen Systemen Inkonsistenzen, Redundanzen und Ambiguitäten anzutreffen sind. Auf der anderen Seite finden sich zunehmend Systeme mit einer reichhaltigen und formal fundierten Konzeptbeschreibungssprache wie etwa GRAIL oder K-REP. Die formalsprachlichen Ausdrucksmittel nutzend weisen Wissensbasen dieser Klasse sehr ausdifferenzierte Konzeptbeschreibungen auf, allerdings decken sie nur stark beschränkte Ausschnitte der medizinischen Domäne ab. Der wissenschaftliche Konsens entwickelt sich dahin, daß einerseits große Wissensrepositorien ohne eine formale Fundierung nicht auskommen (wie z.B. durch die Entwicklung von SNOMED-RT auf der Grundlage eines beschreibungslogischen Wissensrepräsentationssystems verdeutlicht (Spackman

& Campbell, 1998)). Andererseits sind die sprachlich mächtigeren Systeme wegen ihrer derzeit unzureichenden konzeptuellen Abdeckung hochzuskalieren.

Durch den semi-automatischen Import von Wissensrepositorien, die über eine große konzeptuelle Abdeckung bei gleichzeitiger semantischer Unterdeterminiertheit verfügen, in formal fundierte Wissensrepräsentationssysteme läßt sich die Lücke zu den Wissensrepositorien mit schmaler Abdeckung und formaler Fundierung schließen. Erste Ansätze hierzu gehen bereits auf Carenini & Moore (1994) zurück. Hierbei wurde versucht, das semantische Netzwerk von UMLS dazu zu verwenden, die dadurch gewinnbaren Konzeptbeschreibungen mithilfe eines graphischen Tools in LOOM zu übertragen. Inzwischen gehen die Arbeiten deutlich weiter, so daß tatsächlich eine große Anzahl von Konzepten in einer Wissensrepräsentationssprache zur Verfügung steht (z.B. Pisanelli et al. (1998) und Schulz et al. (1999)). Bedingt durch die in UMLS vorhandenen Fehler kann der Wissensimport freilich nur teilweise automatisiert durchgeführt werden. Neben der Generierung von Konzept- und Relationsdefinitionen in der jeweiligen Beschreibungssprache müssen redundante Definitionen, terminologische Zyklen in der generischen Hierarchie sowie die Vermischung von Generalisierungsbeziehungen mit partitiven Beziehungen manuell bereinigt werden. Schließlich sind Lücken im Wissen durch manuelle Verfeinerungen zu schließen (Schulz & Hahn, 2000).

Auch für die Skalierung des in der Medizin auf die Teildomäne der gastro-intestinalen Erkrankungen beschränkten Textverstehenssystems MEDSYNDIKATE spielt die Erweiterung des zur Verfügung stehenden Domänenwissens eine große Bedeutung. Angesichts der schieren Menge von medizinischen Konzepten erweist sich der Weg des semiautomatischen Wissensimports als wichtige Alternative zur rein manuellen Wissensmodellierung. Das dadurch in großem Umfang gewinnbare Wissen kommt dem Textverstehen unmittelbar zugute, weil erst hierdurch eine ausreichende konzeptuelle Abdeckung erreichbar ist.

Einen groben tabellarischen Überblick über die Verwendbarkeit medizinischer Wissensrepositorien für die Analyse natürlicher Sprache –allerdings ohne jegliche empirische Fundierung– gibt Baud et al. (1996) bezüglich vier Dimensionen: abgedeckte natürliche Sprache(n), erfaßte Terme bzw. Konzepte, verwendete epistemische Primitive zur Kodierung von Term- bzw. Konzeptbeziehungen sowie Ausdrucksstärke des zugrundeliegenden Wissensrepräsentationssystems. Eine detaillierte Erörterung, wie begrifflich-konzeptuelles Wissen auf natürliche Sprache abzubilden ist und wie konzeptuelles Wissen während des Textverstehens zur Anwendung kommen kann, findet sich in Rassinoux et al. (1998).

Kapitel 6

Evaluierung der semantischen Komponente von SYNDIKATE

Nachdem in den vorangegangenen Kapiteln ausführlich der Kontext, in dem die semantische Interpretation stattfindet, aufgezeigt wurde, beschäftigt sich dieses Kapitel mit der Evaluation der semantischen Interpretation. Dabei kommen ausschließlich aus Zeitschriften- bzw. Klinikumstexten unserer beiden Domänen, IT und Medizin, extrahierte Phänomene zur Analyse.

Zwar existiert eine Reihe wichtiger wissenschaftlicher Arbeiten, die sich mit der Bewertung der Wissensextraktionsleistungen von Textverstehenssystemen beschäftigen. Jedoch stellt die Evaluierung der semantischen Interpretation in SYNDIKATE wesentliche neuartige methodische Probleme, die in dieser Form noch nicht angegangen wurden. Insgesamt waren im Zuge dieser Evaluation fast 2.500 Instantiierungen der semantischen Interpretationsfunktion, verteilt auf die drei betrachteten Phänomenklassen, Genitivattribute, Modal- und Auxiliarkonstruktionen sowie Präpositionalphrasen, zu bewerten. Andere Sprachphänomene, die ebenfalls durch die semantische Interpretation abgedeckt sind, kamen nicht in die Evaluation, weil in der Wissensbasis keine ausreichenden Spezifikationen vorliegen (dies betrifft beispielsweise die Interpretation von Adjektiven). Um die Probleme bei der Bewertung ermessen zu können, sei noch einmal daran erinnert, daß SYNDIKATE aus einer Menge von Texten ein Maximum an Fakten zu extrahieren versucht und somit zunächst einmal jede Phrase unabhängig von ihrer Informativität zu interpretieren ist (vgl. hierzu die Ausführungen in Kap. 2). Die Verstehensleistung ist daher nicht an prädefinierte Templates gebunden, sondern orientiert sich am vorgehaltenen Weltwissen und ist dabei in der

Lage, komplexe Interpretationsmuster in Form von Relationspfaden zu berechnen. Dieses tiefe Verstehen setzt auf einer feinkörnigen Modellierung der jeweils betrachteten Teil-domänen auf. Damit determiniert das vorhandene Weltwissen, welche Inhalte überhaupt aus einem Text in den Textwissensbasen enthalten sein können. Das hiermit verbundene Knowledge-Engineering-Problem wurde bereits im vorangegangenen Kapitel 5 behandelt.

Bevor aber die eigentliche Evaluation in Angriff genommen werden kann, gilt es zunächst, eine saubere Datengrundlage zu schaffen. Dazu wurden zunächst die der Evaluation zugrundeliegenden Texte in einem randomisierten Verfahren aus den Korpora ausgewählt (vgl. hierzu Kapitel 6.2). Diese Randomisierung ist wichtig für die Verallgemeinerbarkeit der Aussagen. Ein weiteres wichtiges Anliegen dieser Arbeit ist es, zunächst einmal *alle* in den Texten vorkommenden Okkurrenzen der behandelten Sprachphänomene zu extrahieren, um dann deutlich zu machen, für welche Phänomenklassen keine Lösungen vorliegen bzw. andere Lösungsstrategien greifen. Auch die wegen Beschränkungen bei der Verstehensleistung notwendigen Selektionen aus dem Datenmaterial sind transparent gemacht und teilweise durch Klassifikationsaufgaben abgesichert (Kapitel 6.5.2.1). Dies geschieht in ganz bewußter Absetzung zu Arbeiten wie etwa der von Gomez et al. (1997), der in seine empirische Arbeit nur diejenigen sprachlichen Phänomene einbezieht, welche durch die von ihm implementierte semantische Interpretation abgedeckt sind.

Insgesamt steht die Transparenz der Evaluationsverfahren an erster Stelle, wozu auch gehört, daß die angewandten Methoden diskutiert und deren Schwächen deutlich angesprochen werden (Kapitel 6.1 und Kapitel 6.3).

Erst im Anschluß an die Extraktion einer Datengrundlage und die Diskussion der Verfahren zur Evaluation erfolgt der eigentliche Darstellungsteil, der die Ergebnisse präsentiert und diskutiert. In der Evaluation geht es einerseits darum, die beiden Interpretationskonstellationen *direct* und *mediated linkage* abzudecken, andererseits stellen insbesondere die mit in die Evaluation aufgenommenen hochfrequenten und vielschichtigen Präpositionalphrasen den Ansatz zur semantischen Evaluation auf die Probe. Die Ergebnisse zur Evaluation finden sich in den Kapiteln 6.4, 6.5.1 und 6.5.2.

6.1 Methoden zur Evaluierung der Interpretationsleistungen in Textverstehenssystemen

Prinzipiell gilt es, bei der Evaluierung von Systemen, die Wissen aus Texten extrahieren, das Verfahren zu objektivieren und subjektive Momente bei der Bewertung möglichst auszuschließen. Dabei sind hauptsächlich drei Ebenen berührt: die Entwicklung des Evaluationsszenarios, die Systemläufe zur Ergebnisgewinnung sowie schließlich die Bewertung der Ergebnisse. Ferner muß dafür gesorgt werden, daß die Evaluation auf einer sauberen Datengrundlage geschieht, da sonst gewonnene Ergebnisse nicht verallgemeinerbar sind. Außerdem kann jedwedes System für eine spezielle Aufgabe optimiert werden, so daß die Frage bestehen bleibt, ob die allgemeine Performanz sich tatsächlich an der in der empirischen Studie gewonnenen ablesen läßt.

Um eine Verzerrung der Evaluationsresultate durch subjektive Voreingenommenheit zu vermeiden, schlagen Friedman & Hripcsak (1998) unter Berücksichtigung der Erfahrungen aus den MUCs einen kompletten Maßnahmenkatalog vor. Von den zur Objektivierung des Evaluationsverfahrens geforderten Kriterien erfüllt die im folgenden aufgezeigte Evaluation der semantischen Interpretation gerade zwei:

- Den Entwicklern des Systems sind die Testdaten unbekannt.
- Vor der Analyse der Testdaten wurde das System eingefroren, d.h., keinerlei Modifikationen waren zulässig.

Als nicht erfüllbar erwiesen sich indessen die vier übrigen Kriterien:

- Die Entwickler des Systems sollten die Evaluation nicht selbst durchführen.
- Wenn es darum geht, die Verallgemeinerbarkeit zu testen, dann sollten die Entwickler des Systems keinerlei Details über die Studie zur Evaluation kennen.
- Das vollständige Design der Evaluation sollte nicht von den Entwicklern des Systems durchgeführt werden.
- Wenn Domänenexperten eingesetzt werden, um den Referenzstandard zu etablieren, dann sollten sie nicht Entwickler des Systems sein.

Daß sich diese vier Kriterien nicht erfüllen ließen, ist im wesentlichen auf zwei Faktoren zurückzuführen. Zum einen existiert für deutschsprachige Texte kein Goldstandard in Form von semantischen Annotationen, mit dem die Performanz eines Textverstehenssystems gemessen werden könnte. Zum anderen gibt es für deutschsprachige Texte keine professionell verwaltete Projektkontexte vergleichbar den MUCs, die sich neben der Klärung methodischer Fragen die Objektivierung des Evaluationsverfahrens zum Ziel gesetzt hat, indem beispielsweise auch semantisch annotierte Textkorpora bereitgestellt werden.

Aus diesem Grund folgt die Darstellung im folgenden der weiteren Empfehlung von Friedman & Hripcsak (1998), das Verfahren zur Evaluation offenzulegen und die in ihm liegenden Schwächen zur Darstellung zu bringen. Immerhin können die nachfolgenden Kapitel, die das Verfahren und die Ergebnisse der Evaluation präsentieren, die wesentlichen Forderungen von Friedman & Hripcsak (1998) erfüllen, wenn es um die Präsentation der Ergebnisse geht. Die Bewertung der Ergebnisse zur semantischen Interpretation gibt die Aussagekraft der Übereinstimmung zwischen den Klassifizierenden (Interrater Reliability) mit Hilfe der κ -Statistik an (Siegel & Castellan, 1988). Da die Testtexte in einem randomisierten Verfahren aus dem Textkorpus gezogen wurden, können aufgrund der dadurch erzielten Verallgemeinbarkeit die jeweiligen Konfidenzintervalle angegeben werden. Schließlich sind die Formeln, aus denen sich die Performanzwerte berechnen, ausführlich beschrieben.

6.2 Extraktion der Datengrundlage aus den Textkorpora

Eine mittlerweile leicht zu gewährleistende Grundvoraussetzung bei der Evaluation besteht darin, eine ausreichende Menge von Texten in maschinenlesbarer Form für jede der von SYNDIKATE behandelten Domänen, IT und Medizin, zur Verfügung zu haben. In der Medizin fallen im Krankenhausalltag Unmengen an elektronisch gespeicherten Freitexten an. Allerdings liegen sie oftmals in einer medizininformatisch vorgegebenen Struktur vor, die die unmittelbare Verwendung ausschließt.¹ Geeignete Transformationsmechanismen vor-

¹So war es in der empirischen Studie ursprünglich geplant, Daten aus dem Universitätsklinikum in Graz, Österreich, einzusetzen, um die Datengrundlage weiter zu diversifizieren. Bedauerlicherweise lagen die Histologieberichte als Datensätze in strukturierter Form, also nicht als Freitext, vor und konnten daher

ausgesetzt, um Texte in ein weiterverarbeitbares Format zu bringen, bietet die Medizin ein nahezu unerschöpfliches Reservoir an Textmaterial. Um in der IT an eine ausreichende Menge an Texten zu gelangen, wurden etliche Redaktionen von in der Bundesrepublik erscheinenden Zeitschriften über Computer und verwandte Themen angeschrieben und um die Überlassung von maschinenlesbaren Textdateien gebeten. Insgesamt konnten auf diese Weise 16,5 Jahrgänge von fünf verschiedenen Computerzeitschriften mit unterschiedlichem inhaltlichem Niveau akquiriert werden.² Die Anzahl aller in diesem Korpus enthaltenen Artikel beläuft sich auf ca. 50.000.

Auf der systeminhärenten Ebene beeinflussen die Evaluation folgende Faktoren: bestehende Spezifikationslücken auf allen Wissens Ebenen, der thematische Fokus in der Informationstechnologie auf das Schwerpunktthema Hardware sowie der thematische Fokus in der Medizin durch die Beschränkung auf gastrointestinale Erkrankungen und schließlich die Eingrenzung auf die Textsorten Produktberichte (IT) und Histologieberichte (Med). Aus diesen Einschränkungen heraus folgt – allerdings mit je unterschiedlicher Schärfe –, daß bei der Textselektion nicht alle in den Textkorpora vorhandenen Texte auch tatsächlich für eine aussagekräftige Evaluation verwendet werden müssen. Die Verfahren, mit denen die zur Evaluation herangezogenen Texte ermittelt wurden, sollen nun im folgenden für die Informationstechnologie wie für die Medizin aufgezeigt werden.

6.2.1 Extraktion der Datengrundlage aus IT-Texten

Aus den 50.000 IT-Texten wurde eine einprozentige randomisierte Stichprobe von 500 Texten gezogen. Da hierbei keinerlei Vorbeschränkungen für die Auswahl gegeben waren und die gewählten Zeitschriften zugleich eine enorme inhaltliche Breite aufweisen, enthielt die Stichprobe eine stattliche Anzahl von Texten, die deutlich außerhalb des von SYNDIKATE abgedeckten thematischen Skopus' liegen. Aus diesem Grund war eine empirische Vorstudie zur Ermittlung der inhaltlich analysierbaren Texte unerläßlich, weil damit die in den Texten vorhandenen Sprachphänomene größtenteils innerhalb der vorhandenen Spezifikationen liegen sollten. Die Erweiterung der Spezifikationen hätte wegen des Fehlens verfügbarer größerer Wissensressourcen einen nicht realisierbaren Skalierungsaufwand erfordert.

leider keine Verwendung finden.

²Im einzelnen handelt es sich um *Computer und Technik (ct)*, *PC Professional*, *PC Magazin*, *PC Welt* und *PC Direkt*.

Innerhalb dieser empirischen Studie zur Klassifikation der ausgewählten Texte waren insgesamt acht Juroren zwei Merkmalsdimensionen vorgegeben, anhand derer die Texte einzustufen waren.³ Die erste Dimension betraf die Textsorte. Die Textsorte stellt deshalb eine wichtige Beschreibungsebene dar, weil die Verstehensleistung durch sie beeinflusst wird. Ein fiktionaler Text enthält zum Beispiel i.a. eine Reihe von Stilmitteln (Ironie, eingeschobene Dialoge etc.), die von keinem Textverstehenssystem vollständig abgedeckt werden. Auch lassen sich fiktionale Texte thematisch nur schwer einordnen, weshalb deren Bezug zur Informationstechnologie sehr weitläufig sein kann. Dennoch sind fiktionale Texte in unserem IT-Textkorpus enthalten und mußten deshalb ausgefiltert werden.

Die zweite erfaßte Dimension betrifft das Thema eines Texts. Auch hier zeigt sich eine erstaunliche Variabilität innerhalb des Textkorpus'. Eine größere Anzahl der Texte setzt sich mit juristischen Fragen auseinander, etwa der Gültigkeit von Allgemeinen Geschäftsbedingungen bei Geschäften über das Internet, oder nahm Bezug auf wirtschaftliche Daten, wie Quartalsumsätze, Umsatzrenditen oder Firmenaufkäufe. Solche Texte fallen letztlich vollständig aus dem aktuellen Skopus von ITSYNDIKATE hinaus, obwohl sie in IT-Zeitschriften anzutreffen sind.

In der eigentlichen Klassifikationsaufgabe mußten die 8 Klassifizierer jedem der Texte, der in der Stichprobe enthalten war, für jede der beiden Dimension *genau* eine Kategorie zuordnen. Zu jeder Kategorie wurde eine Definition in Form einer natürlichsprachlichen Beschreibung gegeben.⁴ Die inhaltliche Füllung der Kategorien innerhalb der Dimensionen erfolgte auf der Grundlage der in den Computerzeitschriften anzutreffenden Strukturierungen, die aus den jeweiligen Inhaltsverzeichnissen abgeleitet sind; diese wurden allerdings entsprechend den Anforderungen der Klassifikationsaufgabe angereichert. Als wichtiges Kriterium bei der Kategorienbildung galt vor allem, daß keine Interaktion zwischen den Dimensionen stattfindet. D.h., die Auswahl einer Kategorie in der Dimension *Textsorte* determiniert in keiner Weise, welche Kategorie in der Dimension *Thema* auszuwählen ist. Zur Dimension Textsorte gehören sieben Kategorien und zur Dimension Thema insgesamt neun (eine genaue Erläuterung enthält Anhang A). Aufgrund der Unabhängigkeit bei der Kombination ergab sich damit insgesamt eine Auswahl aus 63 Möglichkeiten.

Da sich ITSYNDIKATE auf Produktberichte über Hardware konzentriert, sind in die Evaluation der semantischen Interpretation nur diejenigen Texte aufgenommen worden,

³Eine detaillierte Beschreibung des Verfahrens findet sich in Anhang A.

⁴Anhang A gibt diese Definitionen wieder.

Treffer	Binomialverteilung
≥ 0	1.
≥ 1	0.12
≥ 2	0.01
≥ 3	$2.11 * 10^{-4}$
≥ 4	$4.22 * 10^{-6}$
≥ 5	$5.42 * 10^{-8}$
≥ 6	$4.36 * 10^{-10}$
≥ 7	$2.00 * 10^{-12}$
8	$4.03 * 10^{-15}$

Tabelle 6.1: Wahrscheinlichkeitsverteilung bei der Textklassifikation

die sowohl Textsorte *Bericht* als auch zum Thema *Hardware* gehören. Das festgelegte Einschlußkriterium, durch dessen Anwendung ein Text aus dem Korpus selektiert wurde, bestand aus der Übereinstimmung von sieben der acht Klassifizierer in der einem Text zugeordneten Dimension. Mit Hilfe der kumulativen Wahrscheinlichkeitsverteilung für Binomialtest läßt sich die Signifikanzgrenze ermitteln, die eine gute Übereinstimmung der Rater angibt (Hartung et al., 1991). Grundlage dieses Verfahrens ist die Berechnung der Wahrscheinlichkeit, daß alle Rater zufällig übereinstimmen. In Tabelle 6.1 sind die Werte für die Wahrscheinlichkeitsverteilung einer zufälligen Übereinstimmung von einem bis maximal acht Ratern angegeben. Die Signifikanzgrenze von $p \leq 0.01$ wird bereits erreicht, wenn mehr als drei der Klassifizierenden übereinstimmen (siehe Tabelle 6.1). Aus Tabelle 6.1 wird auch ersichtlich, daß die Festlegung auf mindestens sieben Rater mit gleichem Ergebnis ein ziemlich hartes Inklusionskriterium darstellt.

Durch die Beschränkung auf Berichte über Hardware – die mit Hilfe der Klassifikationsaufgabe umgesetzt wurde – reduzierte sich das Textsample von 500 auf 48 Texte, die allerdings nach wie vor das gesamte Spektrum der im Korpus anzutreffenden Zeitschriften reflektieren. In einem Folgeschritt wurde das Textsample randomisiert in etwa im Verhältnis 2 : 1 geteilt. 33 dieser Texte dienten als Trainingsset. Die Funktion des Trainingssets besteht darin, zu ermitteln wie gut die Verstehensleistung des Systems bei optimaler manueller Adjustierung sein kann. Damit ist es möglich, auf dieser Ebene einen – wenn auch nur schwachen – Goldstandard bezüglich einer oberen Schranke dafür zu fixieren, welche Per-

formanz das System beim aktuellen Stand der Implementation erreichen kann. Das Testset kam in einem zweiten Systemlauf zum Einsatz, um zu evaluieren, welche Ergebnisse die semantische Interpretation bei unbekanntem Textinhalt zu liefern in der Lage ist.

Der Begriff der Trainingssets muß auch dahingehend präzisiert werden, daß auch Wissen im System implementiert ist, das über das zum Analysieren ausschließlich dieser Trainings-texte notwendige Wissen deutlich hinausgeht. D.h., neben dem erwähnten Zweck, die obere Grenze für die Verstehensleistung zu bestimmen, dient das Trainingsset auch dazu, zu ermitteln, welche Konzepte für das Verstehen in der ausgewählten Domäne essentiell (d.h. hochfrequent) sind, um dadurch den Modellierungsaufwand in der IT-Wissensbasis eingrenzen zu können. Die restlichen 15 Texte dienen als Testset für den zweiten Teil der Evaluation.⁵

Die 33 Texte des Trainingssets mit insgesamt 6.732 Wörtern enthalten durchschnittlich 204 Wörter (siehe Tabelle 6.2). Der kürzeste Text kam auf nur 43 Wörter, der längste auf 2.343 Wörter. Die Standardabweichung liegt bei 397. Allerdings stellt der längste Text einen ziemlichen “Ausreißer” dar, da der zweitlängste Text auf nur 597 Wörter kam. Die Angabe des Medians von 102 Wörtern vermag dies noch weiter zu verdeutlichen. In den 15 Texten des Testsets finden sich im Durchschnitt 144 Wörter bei 2.159 Wörtern insgesamt. Mit 75 Wörtern ist der kürzeste Text des Testsets noch länger als insgesamt vier Texte des Trainingssets. Der längste Text liegt mit 684 Wörtern in etwa gleich auf mit dem zweitlängsten Text im Trainingsset. Für die Testtexte ergibt sich daher eine im Vergleich zum Trainingsset niedrigere Standardabweichung von 152. Der Median liegt bei 101. Offensichtlich haben die Vorgaben, nach denen die Texte zu klassifizieren waren, dazu geführt, daß mehrheitlich auffällig kurze Produktbeschreibungen in das Trainings- bzw. das Testset aufgenommen wurden.

6.2.2 Extraktion der Datengrundlage aus medizinischen Texten

Innerhalb der Medizin gestaltete sich die Extraktion eines Textsamples nicht ganz so schwierig, da durch die Provenienz der Texte aus dem Pathologischen Institut des Uni-

⁵Die nicht ganz dem ursprünglich anvisierten Teilungsverhältnis entsprechende Aufteilung von 33 : 15 resultiert aus dem Umstand, daß im Trainingsset zwei der Textsamples eigentlich aus zwei bzw. drei unterschiedlichen Texten bestehen, auch wenn sie in den elektronischen Publikationen in einem File zusammengefaßt sind. Die nachträgliche Korrektur erhöhte die Anzahl der Texte dann von 30 auf 33.

	IT-Training	IT-Test	MED-Training	MED-Test
Anzahl Texte:	33	15	59	30
Anzahl Wörter:	6.732	2.159	7.754	3.567
Mittel:	204, 0	143, 9	131, 4	118, 9
Median:	102, 0	101, 0	112, 0	122, 5
Standardabweichung:	396, 8	152, 2	67, 4	51, 1

Tabelle 6.2: Statistische Auswertung zur Wortanzahl in IT- und Medizin-Texten

versitätsklinikums in Freiburg (Zaiß et al., 1991) sowohl die Textsorte (Histologiebericht) als auch das Thema (alle Bereiche der Pathologie mit Ausnahme der Neuropathologie) bereits festliegen. Im folgenden wurden zunächst sämtliche histologische Gutachten des 1. Quartals des Jahres 1999 als Grundlage zur weiteren Bearbeitung ausgewählt. Diese summieren sich auf exakt 4.973 Befundberichte. Das Einschlußkriterium für die Aufnahme eines Texts in die Kollektion der relevanten Texte war wegen der feststehenden Textsorte ein rein inhaltliches: Befunde zu Schleimhautbiopsien des oberen Verdauungstrakts. Eine zur Vereinfachung des Verfahrens nur mit Hilfe einer Menge von Suchtermen ermittelte Anzahl inhaltlich relevanter Dokumente wurde manuell durchgesehen und die 159 falsch-positiven Dokumente daraus beseitigt. Auch die Restmenge von Texten, die nach dem ersten Durchlauf nicht zu den relevanten Dokumenten gehörten, wurde einer manuellen Durchsicht unterzogen, so daß nochmals 17 Falschnegative aufgespürt werden konnten. Insgesamt verblieben damit 752 relevante Texte, aus denen randomisiert die zur Evaluati-on dienende Stichprobe gezogen wurde. Wegen der geringeren inhaltlichen Varianz bei den histologischen Befundberichten ist die Anzahl der in der Textkollektion für die medizinische Domäne enthaltenen Texte mit 89 für Training und Test etwas höher als für die IT, und zwar bei ungefähr gleichem Verhältnis (59:30).⁶

Die 59 Texte des Trainingssets enthalten 7.754 Wörter, was einen Durchschnitt von 131,4 ergibt. Der Median liegt bei 112 Wörtern. Die Standardabweichung beträgt 67. Der kürzeste Text weist 52 Wörter auf, der längste Text hingegen 429. In den 30 Texten des Testsets finden sich insgesamt 3.569 Wörter, im Durchschnitt also 119 Wörter. Der Median

⁶Die Datenbank, aus der die Texte stammen, enthielt an einer Stelle eine Doublette. Dies fiel allerdings erst bei genauerer Analyse der Texte auf. Aus diesem Grund finden sich statt 60 nur 59 Texte im Trainingsset.

liegt bei 123 Wörtern. Die Standardabweichung beträgt 51. Die Textlänge reicht von 45 Wörter im kürzesten Text bis 298 im längsten Text (vgl. zu allen vorgenannten Zahlen Tabelle 6.2). Insgesamt betrachtet sind Befundberichte kurze und kompakte Texte mit einer geringen Schwankungsbreite bei der Textlänge. In Anhang C findet sich je ein Beispieltext aus der Informationstechnologie und Medizin.

6.3 Evaluierungsszenario für die semantische Interpretation in SYNDIKATE

Im Zusammenhang mit der Evaluation der Verstehensleistung, die Systeme zur Verarbeitung natürlichsprachlicher Texte bzw. Teilkomponenten von ihnen ermöglichen, existieren eine Reihe von begrenzenden Faktoren, die teilweise systeminhärent sind, teilweise jedoch auch systemextern anzusiedeln sind. Zu den externen Faktoren gehört insbesondere, zumindest für Systeme mit deutschsprachigen Eingabetexten, das Fehlen eines Goldstandards in Form von semantisch annotierten Korpora, gegen die die Verstehensleistung gemessen werden könnte. Der Aufbau und die Pflege derartiger *“Bänke für semantische Interpretation”* ist zeitaufwendig, kostspielig und kann nur von speziell geschultem Personal durchgeführt werden. Andererseits gibt es längerfristig keine methodische Alternative zu diesem auf Normierung und Objektivierung ausgerichteten Ansatz, einen Goldstandard zu schaffen, mit dessen Hilfe die Qualität der inhaltlichen Interpretation eines Texts zu beurteilen ist. Da der Evaluation von SYNDIKATE kein derartiger annotierter Korpus zur Verfügung steht, basiert die Beurteilung der semantischen Interpretation auf einem introspektiven Urteil. Jedoch erweist sich in diesem Kontext die tiefe und feinkörnige Modellierung des Domänenwissens als ausgesprochen nützlich, um den subjektiven Faktor bei der Introspektion gering zu halten, denn sowohl Relationen wie auch Werterestriktionen für den Konzepttypen eines Relationsfüllers sind scharf voneinander abgegrenzt.

Natürlich stellt sich im Zusammenhang eines komplexen Textverstehenssystems die Frage, was in dieser empirischen Arbeit zur Evaluation des Teilmoduls der semantischen Interpretation nun tatsächlich evaluiert wird. Um Fehler aus anderen Modulen auszuschließen, gehen wir von idealisierten Randbedingungen aus. Unter diesen Vorannahmen instantiiert das Lexikon die richtigen Konzeptidentifizier, auch wenn unter Umständen mehrere Wörter zu einer Bedeutung (einer Menge von Konzeptidentifiern) zusammengefaßt sind, wie dies

zum Beispiel bei Funktionsverben der Fall ist (*zur Verfügung stehen* $\Rightarrow$ {BE-AVAILABLE}). Der Parser ist in der Lage, die richtigen Abhängigkeitsrelationen zu berechnen, die zu einem Triggern der semantischen Interpretation führen. Um bei Präpositionalphrasen die Menge der Anfragen zu begrenzen, sind nur semantisch gültige Instantiierungen der Interpretationsschemata vorgenommen worden. Was in dieser empirischen Studie *nicht* getestet wurde, ist die Sensitivität des Interpretationsverfahrens für das *Semantic Filtering*, d.h., die Fähigkeit, syntaktisch erlaubte, aber semantisch-konzeptuell zu verwerfende Anfragen auch mit einem negativen Ergebnis zu beantworten. In demselben Zusammenhang kam auf der Ebene der Ambiguitäten, was insbesondere wiederum die Präpositionalphrasen betrifft, eine äußerst restriktive Strategie zum Tragen. In einem Konsensverfahren mußten sich fünf Rater, die die Präpositionalphrasen in den Texten zu klassifizieren hatten (s.u.), auf eine eindeutige Menge von Anfragen einigen, um eine einheitliche Datengrundlage zu schaffen. Hierunter fällt auch, zu entscheiden, in welchen Fällen eine echte Ambiguität unter Einbeziehung aller Wissensquellen vorliegt. Nur die eindeutigen Fälle kamen in die eigentlichen Trainings- und Testsets. Folglich bestehen die verwendeten Sets jeweils aus einer Menge von Instantiierungen der jeweilig anzuwendenden semantischen Interpretationsfunktion mit den aktuellen Parametern, die der dazugehörige minimale interpretierbare Subgraph im Abhängigkeitsgraph liefert.

Im Prinzip ließe sich die Evaluierung der semantischen Interpretation in zwei (bzw. sogar drei) Teile zergliedern. Zum einen wäre zu prüfen, wie gut die statischen Constraints, die an den Wortklassen und Abhängigkeitsrelationen bzw. an den intervenierenden lexikalischen Knoten hängen, in der Lage sind, tatsächlich die Interpretation zu restringieren und ob nicht vielleicht zu viele semantische Interpretationen durch Überrestringierung kein Ergebnis liefern. Zum anderen käme eine isolierte Betrachtung des dynamischen Interpretationsprozesses in Betracht, der, gegeben den azyklischen Graphen des Konzeptsystems, valide Pfade berechnet. Schließlich können beide Einzelaufgaben in einem dritten Szenario miteinander verbunden werden. Die hier dargestellte Evaluation beschreibt das kombinierte Szenario.

Die in den nachfolgenden Tabellen 6.3, 6.4 und 6.5 enthaltenen Daten und Wertungen geben einen umfassenden Überblick über die Quantität und Qualität der semantischen Interpretation für die drei Phänomenklassen Genitivattribute, Modal- und Auxiliarkonstruktionen sowie schließlich Präpositionalphrasen bezüglich der Trainings- und Testsets in der IT und Medizin. Die Ergebnisse der semantischen Interpretation der damit insge-

6.3 Evaluierungsszenario für die semantische Interpretation in SYNDIKATE193

samt zwölf Systemläufe wurden jeweils von zwei Ratern bewertet. Für die IT-Texte gaben ein Student, der an der UPPER-MODEL- und IT-Domänen-Modellierung mitgewirkt hatte, und der Autor dieser Dissertation die Wertungen ab, für die medizinischen Texte ein Mediziner, welcher für die Modellierung in der Medizin-Domäne zuständig war, sowie wiederum der Autor. Die in den folgenden Tabellen angegebenen absoluten Zahlen, die die Rating-Ergebnisse wiedergeben, ergeben sich aus dem Mittelwert der Bewertungen, die die beiden Rater getroffen haben. Eine statistische Bewertung der Übereinstimmung beim Rating erfolgt mit Hilfe der κ -Statistik zur Bestimmung der Inter-Rater-Reliability (zu den Vorteilen dieses Verfahrens siehe Carletta (1997)).

Die folgenden Ausführungen beschreiben das methodische Vorgehen, um zu den Wertungen zu gelangen, die zur Qualifizierung der Ergebnisse verwendeten Metriken sowie die sonstige in diesen Tabellen enthaltene Information.

Die eigentliche Bewertungsaufgabe schließt prinzipiell nur die Kategorien *falsch* und *richtig* ein. Eine Bewertung kann allerdings nur dann erfolgen, wenn eine semantische Interpretation überhaupt ein Ergebnis erzeugt. Es gilt zwei Fälle zu unterscheiden, in denen die semantische Ebene kein Ergebnis liefert. Im ersten Fall kann die der jeweiligen Phänomenklasse zugehörige semantische Interpretationsfunktion nicht instantiiert werden, weil mindestens eine der beiden involvierten lexikalischen Oberflächen im aktuellen minimalen interpretierbaren Dependenzgraphen kein konzeptuelles Korrelat besitzt; d.h., es liegt eine Unterspezifikation im Weltwissen auf der Ebene des *Konzeptbestandes* vor. Die Tabellen zur Evaluation der semantischen Interpretation listen diese Konstellation unter “**ohne Interpretation**” auf. Im zweiten Fall sind zwar die beiden Konzeptidentifizierer und damit die für die semantische Interpretation benötigten Typinformationen vorhanden, der Interpretationsprozeß gibt indessen keine gültige Relationierung zwischen diesen beiden zurück, obwohl eine solche zu berechnen wäre (was syntaxseitig dann das Einziehen der angefragten Dependenzrelation fälschlicherweise untersagen würde). Da die semantische Interpretation mit einer Anfrage gestartet wird, die ein positives Ergebnis ergeben sollte, ist die Nicht-Rückgabe von mindestens einem gültigen Relationspfad als Fehler zu werten. Wird trotz erfolgter Anfrage kein Relationspfad gefunden, dann läßt sich auf eine unzureichende, bisweilen sogar falsche *Relationierung zwischen Konzepten* durch Konzeptbeschreibungen schließen, d.h. auf eine Unter- und im Extremfall sogar Fehlspezifikation in der Relationshierarchie und deren Verwendung in Konzeptdefinitionen. In den Bewertungstabellen steht dann für diesen Fall unter “**mit Interpretation**” die Marke “*NIL*”.

Berechnet der semantische Interpretationsprozeß tatsächlich eine nicht-leere Menge von Relationierungen, erfolgt die eigentliche Bewertung. Eine Interpretation ist genau dann richtig, wenn in der Menge der Relationspfade mindestens ein für die Rater akzeptabler Relationspfad enthalten ist, ansonsten wird sie als *falsch* gewertet. Um jedoch zwischen eindeutig richtigen Interpretationen und ambigen Interpretationen unterscheiden zu können, sind die Anzahlen jeweils separat unter “**mit Interpretation**” durch die Zeilen “*richtig (eindeutige Lesart)*” und “*richtig (mehrdeutige Lesart)*” aufgeführt.⁷

Die eigentliche Bewertung der Ergebnisse, die die semantische Interpretation berechnet, stellt ein noch nicht vollständig gelöstes Problem dar, da die Ergebnisse selbst nicht unabhängig von der Modellierungskonzeption zu sehen sind. Damit bleiben sie, auch wenn in vielen Bereichen Quasi-Standards wie thematische Rollen zur Anwendung kommen, ein Stück weit systemspezifisch. Aber gerade die im Weltwissens des Systems realisierte feinkörnige Modellierung hilft, den subjektiven Charakter der Bewertung zu mindern.

Beispielsweise kann das Konzept CPU ausschließlich über die konzeptuelle Relation HAS-CPU als Unterrelation von HAS-PHYSICAL-PART an das Konzept MOTHERBOARD angebunden werden. Damit ist eindeutig bestimmbar, ob eine Interpretation falsch oder richtig ist. Falsch ist sie dann, wenn entweder eine Instanz von CPU über eine andere Relation als HAS-CPU an MOTHERBOARD gebunden wird oder wenn eine Instanz mit einem anderen Konzepttyp als CPU in die Relation HAS-CPU eingetragen wird.

Schwieriger – und eigentlich nur noch von geschulten Ratern zu bewältigen – wird das Beurteilen des Ergebnisses dann, wenn die semantische Interpretation einen Relationspfad zwischen den zu relationierenden Instanzen ermittelt, dessen Länge größer ist als eins. Zur Illustration mag nachfolgendes Beispiel dienen. Zu interpretieren sei das Genitivattribut “*die Preise des Herstellers*”. Im Zuge der semantischen Interpretation wird der Kantenzug PRICE PRICE-OF PRODUCT MANUFACTURED-BY MANUFACTURER ermittelt. Da indessen

⁷Vordergründig mag die Festlegung, daß nur mindestens ein Relationspfad richtig sein muß, als ein schwaches Kriterium anmuten. Mit dieser Festlegung trägt man indessen dem Umstand Rechnung, daß es schwierig ist, saubere Metriken dafür zu bestimmen, wie etwa eine Ergebnismenge zu bewerten wäre, die 4 richtige und 2 falsche Pfade enthielte, wenn diese mit Ergebnissen mit nur einem einzigen Pfad zu vergleichen wäre. Da die Anzahl der richtigen und mehrdeutigen Interpretation nur etwas über 10% der insgesamt richtigen Interpretationen ausmacht (180 von 1780) ist der Einfluß auf das Gesamtergebnis indessen nicht zu groß. Darüber hinaus sind die Interpretationsergebnisse mit mehr als drei gefundenen Pfaden ohnehin mehrheitlich als falsch klassifiziert worden. Insgesamt gibt es in allen vier Sets 56 Interpretationen, die mehr als drei Pfade enthalten. 33 (59%) von diesen wurden als *falsch* klassifiziert.

6.3 Evaluierungsszenario für die semantische Interpretation in SYNDIKATE195

die Tatsache, daß Hersteller Preise festlegen, in dieser Interpretation nicht eingeschlossen ist, liegt in diesem Fall eine falsche Interpretation vor. Und dies, obwohl diese Interpretation eigentlich richtig ist, wenn man die aufgrund der Systemspezifikationen erwartbaren Interpretation zugrundelegt.

In Anlehnung an das Information Retrieval kommen die Evaluationsmetriken *Recall* und *Precision* zur Anwendung. Da es sich indessen bei der semantischen Interpretation um keine klassische IR-Aufgabe handelt, sind die beiden Maße in Anpassung an die gegenüber dem IR etwas differenzierteren Ergebnisdaten definiert. Demnach errechnet sich der Recall aus der Summe aller Fälle mit richtiger Interpretation, unabhängig davon, ob das Ergebnis eindeutig oder mehrdeutig ist, geteilt durch die Anzahl aller Okkurrenzen des jeweiligen Phänomens in den Textsets (siehe 6.1).⁸

$$si - recall = \frac{\#richtig(eindeutig) + \#richtig(mehrdeutig)}{\#Vorkommen} \quad (6.1)$$

Im Gegensatz dazu ermittelt sich die Precision aus der Anzahl der richtigen Interpretationen, unabhängig davon, ob diese eindeutig oder mehrdeutig sind, geteilt durch die Anzahl aller überhaupt erzielten Interpretationen einschließlich *NIL* (siehe 6.2).

$$si - precision := \frac{\#richtig(eindeutig) + \#richtig(mehrdeutig)}{\#mitInterpretation} \quad (6.2)$$

Aus den vorstehenden Berechnungsformeln wird ersichtlich, daß die *NIL*-Fälle die Precision beeinflussen (sie zählen zu *#mitInterpretation*), indem sie diese als Teil des Nenners reduzieren, nicht aber den Recall. Die Berechtigung, die *NIL*-Fälle bei der Berechnung des Recalls nicht zu den erzielten Ergebnissen zu zählen, leitet sich daraus ab, daß eine leere Menge an Relationspfaden keine gültige Interpretation darstellt. Hierdurch fehlt dann genau die für das eigentliche Faktenretrieval notwendige Verknüpfung mit dem Text.

Obwohl natürlich weitestgehend versucht wird, Fehler aus anderen Modulen nicht in die Studie hineinzupropagieren, gibt es eine ganz deutliche, aber auch gewollte Abhängigkeit: die Qualität und die Breite des verwendbaren Weltwissens. Tatsächlich testet das zum Einsatz kommende Verfahren auch die Konsistenz und Vollständigkeit des ontologischen

⁸Da alle in den Texten vorkommenden Okkurrenzen von Genitivattributen, Modal- und Auxiliarkonstruktionen sowie Präpositionalphrasen durch mindestens zwei Personen händisch extrahiert wurden, kann davon ausgegangen werden, daß die Anzahl der Vorkommen vollständig erfaßt wurde.

Wissens. Genau hierin liegt der Grund dafür, in einem Trainingslauf die Modellierung soweit zu optimieren, daß sie letztendlich keine Fehler mehr produziert und nur noch die eigentliche semantische Interpretation zu testen ist. Freilich scheitert diese Optimierung in manchen Bereichen der Ontologie daran, daß aus Gründen fehlender ontologischer Klarheit gar keine Weltwissensstrukturen modelliert werden können, so daß die semantische Interpretation wegen Unterspezifikation nicht erfolgen kann.

Um zu dokumentieren, welcher Art die bestehenden Spezifikationslücken sind, geben die Ergebnistabellen 6.3, 6.4 und 6.5 zusätzlich Aufschluß darüber, auf welcher Wissensebene die Spezifikationslücken anzusiedeln sind. Wie bereits die Bewertungsaufgabe zur semantischen Interpretation beruhen die Einstufungen auf subjektiver Introspektion und wurden nur von einem der Rater vorgenommen. Dennoch lassen sich die Zugehörigkeiten zu den einzelnen Kategorien relativ klar bestimmen. Das Auftreten nicht spezifizierter Konzepte im Trainingsset, das ja erklärtermaßen die Performanz der semantischen Interpretation ausreizen sollen, begründet sich durch die bereits erwähnte Unmöglichkeit, eine Reihe von Konzepten sinnvoll mit den expressiven Mitteln einer Konzeptbeschreibungssprache zu modellieren.

- *Domänenmodell.* Das fehlende Konzept gehört als Spezialwissen zur Ontologie der jeweiligen Wissensbasis, d.h. zur IT bzw. Medizin. Okkurrenzen dieser Klasse im Trainingsset lassen auf die Nicht-Modellierbarkeit des dazugehörigen Konzepts schließen (Bsp.: *“Texturkompression”*, *“Epithelveränderungen”*).
- *Generisches Modell.* Ein Konzept ist dann dem UPPER-MODEL zuzurechnen, wenn es zum Allgemeinwissen gehört und damit grundsätzlich auch jeder Spezialdomäne zugehörig sein sollte. D.h., allgemeingültige Konzepte der Art eines COMPUTER-SYSTEM oder anatomische Konzepte wie HEART oder EYE gehören ins UPPER-MODEL. Auch hier lassen Lücken im Trainingsset auf Modellierungsprobleme schließen. In diesen Bereich fallen u.a. verbale Gefüge mit sehr allgemeiner Bedeutung *“etwas unter sich ausmachen”*, *“unterfordert sein”* sowie Substantive, die eine sehr ungenaue Bedeutung haben (*“Genese”*, *“Ummauerung”*).
- *andere Domänen.* Gehört ein Konzept weder zum generischen Modell noch zu einer der beiden Domänen IT bzw. Medizin, dann ist es einer anderen Domäne zuzurechnen. In diesen Bereich fallen etwa in den IT-Texten auftretende Begriffe, die sich

6.3 Evaluierungsszenario für die semantische Interpretation in SYNDIKATE197

inhaltlich auf die fremde Domäne der Telekommunikation beziehen (Bsp.: “GSM-Netz”, “Telefonanlage”).

- *Quantifikation.* Unterbleibt eine semantische Interpretation aufgrund quantifizierender Ausdrücke, dann ist dies auf zwei verschiedene Unterklassen zurückzuführen. Zum einen können generalisierte quantifizierende Ausdrücke auftreten (“*einige der Ankündigungen*”), die beim gegenwärtigen Stand der Implementation grundsätzlich nicht interpretierbar sind. Zum anderen treten quantifizierende Konzepte auf, denen prinzipiell ein konzeptuelles Korrelat zuzuordnen wäre, jedoch bliebe dessen Semantik völlig unterdeterminiert, weshalb in diesem Fall eine Interpretation unterbleibt (“*Ummengen an Speicherchips*”).
- *Abstrakta.* Zunächst liegen abstrakte Konzepte nicht im Zentrum der Wissensmodellierung in den ausgewählten Anwendungsdomänen, da es in diesen um sehr konkrete, größtenteils materielle oder objektivierbare Konzepte geht. Ontologisch betrachtet sind diese abstrakten Konzepte im oberen Teil der Konzepthierarchie angesiedelt und sehr schwer zu kategorisieren. Da sie sich oftmals als nicht modellierbar erweisen, bleiben auch in den Trainingssets Spezifikationslücken im Konzeptsystem bestehen. In diesen Bereich fallen Wörter wie “*Art*”, “*Chancen*” oder “*Erhabenheiten*”.
- *Raum.* Räumliche Konzepte teilen oftmals Eigenschaften von abstrakten Konzepten, sofern sie nicht im Sinn eindeutiger Eigenschaften wie Lateralität interpretierbar sind. Die Bedeutung – und damit die Interpretation – ermittelt sich kontextabhängig. Innerhalb der Phrase “*in der Tiefe der Schleimhaut*” bezeichnet das Konzept für Tiefe mit Sicherheit etwas anderes als innerhalb der Phrase “*in der Tiefe des Weltalls*”. Semantische Konnotationen dieser Art sind im Konzeptsystem nicht modellierbar.
- *Zeit.* In Analogie zu den räumlichen Konzepten existieren in den Texten sehr unspezifische zeitliche Konzepte, deren Interpretation allein daran scheitert, daß bereits seitens der Wissensmodellierung keine Zielstruktur anzugeben ist. In der Phrase “*in der Anfangszeit des Pentium*” referenziert “*Anfangszeit*” auf ein Zeitintervall, das nur durch sehr genaue Kenntnis des Kontextes (Markteinführung und Marktpräsenz einer Prozessorklasse von Intel) bestimmbar ist.
- *bildliche Sprache.* Generell faßt die Kategorie der bildlichen Sprache eine größere Klasse von Sprachphänomenen zusammen, die jedoch gemeinsam haben, daß das

Denotat eines in einer Phrase verwendeten Ausdrucks nicht mit der tatsächlichen Bedeutung zusammenfällt. Neben den bereits vorgestellten Metonymien fallen unter anderem auch Metaphern in diese Kategorie. Zum Beispiel referenziert der Ausdruck *“der Würfel”* in der Phrase *“der Würfel paßt auf jeden Schreibtisch”* auf einen im Vor-
text eingeführten Drucker. Oftmals greift figurative Sprache Konzepte auf, die nicht nur in der Denotation vom Zielkonzept abweichen, sondern die auch sehr schwer zu modellieren sind (*“Grafikkünstler”* für Grafikkarte oder *“König Athlon”* für die gute Performanz der Athlon-Prozessors von AMD). Um abschließend zu illustrieren, wie sehr figurative Sprache das Verstehen selbst einfacher Sätze beeinflusst, sei folgender Beispielsatz aus dem Testset der IT angeführt: *“Das 18,2-GByte-Flaggschiff geht mit einem KampfpPreis von 2510 Mark ins Rennen”*.

- *Sonstiges*. Sonstiges dient als Kategorie, um die restlichen nicht den vorgenannten Kategorien zuzuordnenden Phänomene aufzunehmen.

6.4 Evaluierung von *Direct-Linkage* Subgraphen

Bei der Evaluierung von *Direct-Linkage* Konstellationen in einem Dependenzgraphen beschränke ich mich auf das Sprachphänomen der Genitivattribute. Auf den ersten Blick mag die Interpretation von Genitivattributen als die einfachste anmuten. Jedoch bringt gerade das Fehlen von aus der Syntax ableitbaren Restriktionen, um einen konzeptuellen Pfad zwischen zwei Konzepten zu berechnen, einen großen Suchraum im Konzeptgraphen mit sich. Folglich kann die Interpretation von Genitivattributen als Testfall für eine konsistente Modellierung sowie eine performante Pfadsuche und eine empirisch begründbare Pfadevaluation angesehen werden. D.h., aufgrund der fehlenden syntaktisch ableitbaren Beschränkungen erfolgt die Interpretation nahezu ausschließlich mit Hilfe konzeptuellen Wissens.

Zieht man die Häufigkeiten von Genitivattributen in den vier Textsets heran, dann finden sich im Trainingsset der IT insgesamt 61 Okkurrenzen, in der Medizin hingegen 168. Das Testset zur IT enthält insgesamt nur 19 Genitivattribute, das zur Medizin immerhin 91 (vgl. Tabelle 6.3). Im Gegensatz zur IT, in der deutlich weniger Genitivattribute überhaupt interpretiert werden konnten, erreichte die Medizin in Trainings- und Testset eine

gleichbleibend hohe Anzahl an Interpretationen (94% bzw. 95%).⁹ Darüber hinaus ergab sich in der Medizin kein Unterschied zwischen dem Trainings- und Test bei Recall und Precision, die mit 88%/93% bzw. 90%/93% einen sehr ordentlichen Wert erreichten. Daß beim Testset keine Verschlechterung von Precision und Recall zu beobachten ist, läßt sich unter anderem dadurch begründen, daß typische Konfigurationen, die sich konzeptuell als PHENOMENON + Genitiv mit ANATOMICAL-LOCATION paraphrasieren ließen, besonders häufig auftreten (vgl. hierzu auch die späteren Ausführungen in Kapitel 6.6).

In der IT fallen die Ergebnisse demgegenüber deutlich ab. Nur 77% bzw. 63% der Okkurrenzen konnten überhaupt interpretiert werden. Der Recall beträgt 75% für das Trainingsset und nur noch 53% für das Testset, bei allerdings sehr kleinem Sample. Dennoch bleibt positiv zu vermerken, daß wenn eine Interpretation erfolgt, diese zumeist ein gutes Ergebnis liefert, wie es sich in den Precision-Werten von 97% bzw. 83% ausdrückt. Daß mit einem Recall von 75% im Trainingsset bereits eine Obergrenze für mögliche semantische Interpretationen erreicht ist, läßt sich im wesentlichen auf drei Faktoren zurückführen. Zunächst gibt es in SYNDIKATE noch keine Implementation, mit denen quantifizierende Ausdrücke interpretierbar sind (z.B. *einige der Ankündigungen*). Weiterhin besitzen eine Reihe von Wörtern keine eigenwertige Bedeutung, sondern leiten diese als quasi-deiktische Ausdrücke aus dem Textkontext ab (vgl. Hirst (1997)) und können auf ganze Textabschnitte verweisen (z.B. *der Nutzen dieser Lösung*, wobei die Lösung zuvor in einem ganzen Absatz beschrieben ist). Schließlich stellen abstrakte Begriffe ein großes Problem dar, weil sie schwerlich innerhalb des Konzeptsystems definierbar sind (z.B. *der Nutzen*).¹⁰ Die genaue Klassifizierungen der Fälle ohne Ergebnis findet sich ebenfalls in Tabelle 6.3.

Beim Rating der Ergebnisse, die die semantische Interpretation erzielte, ergab sich in der IT im Trainingsset eine Abweichung zwischen den Ratern von nur eins, im Testset eine Abweichung von null. In der Medizin verhält es sich mit einer Abweichung von drei für das Trainingsset und von eins für das Testset ähnlich. Entsprechend den in der Tabelle 6.3

⁹Die dazugehörigen Konfidenzintervalle für einen 95% -Vertrauenbereich (Sachs, 1984) finden sich in der nachfolgenden Zeile der Tabelle. Teilweise finden sich in der Tabelle 6.3 Werte mit "halben" Okkurrenzen. Dies kommt daher, daß zwei Rater die Ergebnisse beurteilten und dabei z.T. auseinander lagen.

¹⁰Freilich wäre es möglich, ein Konzept ABSTRACT-OBJECT einzuführen und Begriffe wie "Nutzen" als dessen Kindkonzepte zu definieren. Diese Lösung hätte jedoch den Makel, daß sie aus Gründen der Spezifikationsdichte Ad-Hoc-Definitionen einführt, die aus ontologischer Sicht nicht befriedigen können. Daher ist es besser, die Spezifikationslücken als nicht auflösbar bestehen zu lassen.

	Training-IT	Test-IT	Training-Med	Test-Med
Recall	75%	53%	88%	90%
Precision	97%	83%	93%	93%
# Vorkommen ...	61	19	168	91
... mit Interpretation	47 (77%)	12 (63%)	158 (94%)	86 (95%)
[Konfidenzintervalle]	[62%-87%]	[38%-83%]	[89%-97%]	[90%-98%]
..... richtig (eindeutige Lesart)	35,5 (58%)	7 (37%)	125,5 (75%)	48,5 (53%)
..... richtig (mehrdeutige Lesart)	10 (16%)	3 (16%)	22 (13%)	33 (36%)
..... falsch	0,5	0	6,5	3,5
..... NIL	1	2	4	1
... ohne Interpretation	14 (23%)	7 (37%)	10 (6%)	5 (5%)
..... Domänenmodell	–	–	–	2
..... Generisches Modell	–	1	2	2
..... andere Domänen	–	–	–	–
..... Quantifikation	3 (5%)	–	2	–
..... Abstrakta	7 (11%)	4	5 (3%)	1
..... Raum	–	–	1	–
..... Zeit	1	1	–	–
..... figurative Sprache	3 (5%)	1	–	–
..... Sonstiges	–	–	–	–
# Nominalisierungen ...	18	3	42	23
... mit Interpretation	17	3	41	23
..... richtig (eindeutige Lesart)	8	0	41	23
..... richtig (mehrdeutige Lesart)	9	3	0	0

Tabelle 6.3: Verstehensleistung bei Genitivattributen

enthaltenen Daten macht es keinen Sinn, eine κ Statistik zu berechnen und einen Signifikanztest durchzuführen. In der eigentlichen Bewertung stehen ohnehin nur zwei Kategorien zur Auswahl, nämlich *falsch* oder *richtig*. Die Differenzierung nach *eindeutige Lesart* und *mehrdeutige Lesart* und auch die *NIL*-Interpretationen stellen ein reines Auszählen dar. Bedingt durch das Vorherrschen der *richtig*-Interpretationen ist die Verteilung zwischen den beiden Kategorien sehr ungleich, was in einen hohen Erwartungswert mündet (vgl. Siegel & Castellan (1988)). Dennoch drückt die in absoluten Zahlen vorliegende Abweichung der Rater eine sehr gute Übereinstimmung aus.

Neben der Betrachtung aller Genitivattribute findet sich in Tabelle 6.3 außerdem eine Aufstellung der Interpretationsergebnisse von Verbnominalisierungen. Mit einem Anteil von gut 25% in der IT und einem ebenso großen Anteil in der Medizin scheinen Nominalisierungen eine wichtige Rolle zu spielen. Im Laufe der empirischen Studie mußte aufgrund der gewonnenen Evidenzen die ursprüngliche Hypothese, Genitivattribute hätten keine Restriktionen, auf der Ebene der Verbnominalisierungen verworfen werden (vgl. Romacker et al. (1999)). Statt dessen ist für diese eine Positivliste mit den Constraints $\{subject, direct-object\}$ zu definieren, weil ansonsten zu viele falsch-positive Pfade berechnet werden. Die große Anzahl potentiell falsch-positiver Pfade bei Verbkonzepten resultiert aus der Tatsache, daß mit Verbkonzepten verbundene Zustände, Ereignisse oder Prozesse durch sehr allgemeine Relationen beschreibbar sind – wie beispielsweise eine Menge unterschiedlicher Zeit- und Ortsangaben.

Der vom syntaktischen Kopf abhängige Genitiv beinhaltet entweder das syntaktische *subject* (Genitivus subjectivus) oder das syntaktische *direct-object* (Genitivus objectivus) der Paraphrase des Genitivattributs mit der verbalen Form der Nominalisierung: *Die Berechnung der Bildzeilen durch den Prozessor* $\Rightarrow$ *Der Prozessor berechnet die Bildzeilen*. An dieser Stelle muß erwähnt werden, daß Rollennominalen sich von den eigentlichen Verbnominalisierungen unterscheiden und folglich wie normale Genitivattribute zu interpretieren sind. Unter Rollennominalen versteht man Nominalisierungen, die auf Verben zurückgehen, die jedoch eine Rolle an einem Verbkonzept füllen (vgl. Reddig (1988)). Der „Benutzer“ (USER) ist zum Beispiel der AGENT des Verbkonzepts für „benutzen“ TO-USE.

Eine Ungenauigkeit bei der Interpretation von Genitivattributen bei einigen Fällen von Verbnominalisierungen ergab sich indessen in der IT dadurch, daß im Trainingsset das abhängige Genitivattribut offensichtlich Interpretationen nur im Sinn eines *direct-object*, nicht aber als *subject* zulassen. Auf dieser Ebene scheint die Restriktion auf beide syntak-

tische Relationen noch zu generell. Jedoch lieferte das vorliegende Material keine ausreichenden Evidenzen, die bereits eine Subklassenbildung mit dedizierter Interpretationsregel gerechtfertigt hätten.

Obwohl Genitivattribute – mit Ausnahme der Verbnominalisierungen – keine Restriktionen mit sich bringen, weisen die empirischen Daten eine auffällig geringe Ambiguitätsrate (semantische Interpretation mit multiplen Ergebnissen) bei diesen Genitivattributen auf. Zwar gab es insgesamt zehn mehrdeutige Interpretationen, von diesen entfielen indessen nur eine auf unrestringierte Genitivattribute, wohingegen neun bei Genitivattributen mit Verbnominalisierungen auftraten. Dies ist ein deutlicher Indikator dafür, daß die feinkörnige, auf Relationsdifferenzierung mit harten Typrestriktionen aufbauende Konzeptmodellierung in diesen Fällen ausreichende Trennungsschärfe besitzt, um die Anzahl interpretatorischer Lesarten nicht exorbitant anwachsen zu lassen. Im Gegensatz dazu verfügen die konzeptuellen Korrelate für viele allgemeine Verbkonzepte wie *Bedienung*, *Verbreiterung* oder *Berechnung* über in der Rollenhierarchie sehr weit oben angesiedelte Typconstraints, wodurch sich die hohe Ambiguitätsrate von mehr als 50% begründet. In der medizinischen Domäne zeigt sich darüber hinaus ein überaus interessantes Phänomen. Wie in der Informationstechnologie gibt es eine größere Anzahl von Verbnominalisierungen (insgesamt 42), die in analoger Weise durch eine Restriktion auf *subject* und *direct-object* semantisch interpretiert wurden. In diesen Bereich fallen etwa Phrasen der folgenden Art: *“eine Entzündung der Schleimhaut”* oder *“eine Verbreiterung der Zellkerne”*. Jedoch existieren neben den morphologisch bestimmbaren Verbnominalisierungen Konzepte, die bei der Interpretation zu behandeln sind wie die Nominalisierungen, die jedoch der Form nach eindeutig nicht von Verben abgeleitete Substantive sind. Typischerweise stehen die zu den Wortoberflächen gehörenden Konzepte in einer Subsumptionsbeziehung zu Verbkonzepten.

In der medizinischen Wissensbasis definiert zum Beispiel das Konzept INFLAMMATION, das konzeptuelle Korrelat zu *“Entzündung”*, einen Prozeß, der anatomische Strukturen affiziert. Das von INFLAMMATION subsumierte Konzept GASTRITIS ist eine Entzündung, deren betroffene Struktur der Magen ist oder Teile davon. Ohne Zweifel umschreibt der Begriff der GASTRITIS einen Krankheitsprozeß, der über INFLAMMATION in der taxonomischen Hierarchie darüber hinaus vom Verbkonzept PROCESS subsumiert wird. Damit ist auch GASTRITIS an die Verbkonzepthierarchie gekoppelt. Freilich kann dieser semantische Gehalt nicht an der Wortform festgemacht werden, sondern bestimmt sich aus den konzeptuellen Regularitäten der medizinischen Domäne. Um eine korrekte semantische In-

terpretation zu erzeugen, die nicht übergeneralisiert, indem sie zu viele Relationen bei der Pfadsuche zuläßt, sind dieselben Restriktionen anzuwenden, wie sie bei Verbnominalisierungen greifen. Dieses Vorgehen legitimiert sich aus den Daten, nach denen sich allein im Trainingsset der Medizin für Genitivattribute 31 Fälle (18%) dieser Art belegen lassen. Im Testset finden sich 16 Okkurrenzen (ebenfalls 18%) dieser semantischen Klasse. Entsprechende Beispiele aus der Informationstechnologie sind in den Trainings- und Testdaten nicht aufgetreten. Daraus ergibt sich ein Indiz dafür, daß der dargestellte Zusammenhang ein Charakteristikum der Medizin zu sein scheint.

6.5 Evaluierung von *Mediated-Linkage* Subgraphen

Die Konfiguration von *Mediated-Linkage* beinhaltet Subgraphen im Dependenzgraphen, die mehr als zwei lexikalische Knoten umfassen. Im wesentlichen spalten sich die solche Subgraphen in zwei Klassen auf: Konstruktionen, die mit Hilfe von Modal- und Auxiliärverben gebildet werden, und Präpositionalphrasen. Während bei Präpositionalphrasen die Anzahl von lexikalischen Knoten generell auf drei beschränkt ist (mit Ausnahme der extrem seltenen Circumpositionen), können Fügungen, die Modal- und Auxiliärverben kombinieren, mehr als drei Knoten aufweisen. Im Trainings- und Testset zusammengekommen machen diese Modal/Auxiliar-Konstruktionen allerdings nur einen bescheidenen Anteil aus. Es finden sich insgesamt nur 14 Okkurrenzen (3%).

In den nachfolgenden beiden Unterkapiteln kommen die Interpretationsergebnisse für beide Phänomenklassen zur Darstellung. Da die semantische Interpretation von Subgraphen, die Modal- bzw. Auxiliärverben einschließen, im wesentlichen eine Projektion des syntaktischen Subjekts auf den abhängigen Verbinfinitiv bzw. das abhängige Partizip leisten, sind auf dieser Ebene wenige Interpretationsvarianten anzutreffen: Das syntaktische Subjekt wird entweder im Sinne eines *subject* oder, bei Vorliegen einer Passivkonstruktion, im Sinne eines *direct-object* des abhängigen Verbs interpretiert. Wesentlich komplexer präsentieren sich die Interpretationszusammenhänge bei Präpositionalphrasen, da hier beispielsweise in Abhängigkeit von der Valenz, die von der Präposition besetzt ist, unterschiedliche Interpretationsregeln greifen. Aus diesem Grund mußte für Präpositionalphrasen in einem vorangehenden Schritt erst einmal eine verwendbare Datengrundlage geschaffen werden. Der komplexeren Problemstellung entsprechend nehmen die Erörterungen zu Präpositionalphrasen einen größeren Raum ein.

6.5.1 Evaluierung von Modal- und Auxiliarkonstruktionen

Die Beurteilung der Interpretationsergebnisse bei Modal- bzw. Auxiliarkonstruktionen gestaltet sich deshalb besonders einfach, weil die durch die Dependenzrelation eingeführten Wohlgeformtheitsbedingungen an den Verbkonzepten im allgemeinen nur eine einzige Relation, ansonsten höchstens zwei Relationen, zulassen. Nur in sehr wenigen Fällen kommt es daher zu tatsächlichen Ambiguitäten. In diesem Zusammenhang sei jedoch nochmals angemerkt, daß manche Verben sehr allgemeine Constraints für Typen an Rollenfüllern festlegen. Beispielsweise verfügt das Konzept PROVIDE-EVIDENCE des abhängigen Verbs *“nachgewiesen”* im Satz 16 mit dem Konzept OBJECT über keine aussagekräftigen Beschränkungen für die PROVIDE-EVIDENCE-PATIENT Rolle, auf die das Konzept des syntaktischen Subjekts *“Helicobacter-pylori”* abgebildet wird. Im Prinzip ist fast alles nachweisbar, entscheidend für Constraints ist hier der weitere Kontext. Genau dieser Kontext wird durch die hier allerdings nicht weiter empirisch untersuchten Regeln zur konzeptuellen Interpretation miteinbezogen, indem diese Regeln Abhängigkeiten *zwischen* Rollenfüllern am Verb-Konzept ebenfalls überprüfen.

16. Im Oberflächenschleim konnten *Helicobacter-pylori* nachgewiesen werden.

Die der Einfachheit dieses Interpretationsschritts entsprechenden guten Ergebnisse lassen sich der Tabelle 6.4 entnehmen. In der IT liegt der Recall mit 84% bei Trainings- und 85% Testset auf ungefähr gleichem Niveau, im Testset erhöhte sich die Precision gegenüber dem Trainingsset sogar von 96% auf 98%. Wenn eine Modal- bzw. Auxiliarkonstruktion interpretierbar ist, dann ist das Ergebnis fast immer richtig. In der Medizin verhielt es sich ähnlich, wobei das Testset gegenüber dem Trainingsset sowohl im Recall als auch in der Precision etwas stärker abfällt. Im Trainingsset konnte ein Recall von immerhin 94% bei einer Precision von 98% erreicht werden. Im Testset fiel demgegenüber der Recall auf 79% ab und auch die Precision verschlechterte sich auf 84%. Verantwortlich hierfür zeichnen sich im Testset auftretende Sätze beispielsweise der Art:

17. Eine Epitheldysplasie ist erfaßt.

Da laut konzeptueller Constraints nur physikalische Objekt (PHYSICAL-OBJECT) erfaßt werden können, ist eine Dysplasie (DYSPLASIA), welche ein pathologisches Phänomen

	Training-IT	Test-IT	Training-Med	Test-Med
Recall	84%	85%	94%	79%
Precision	96%	98%	98%	84%
# Vorkommen ...	167	53	131	55
... mit Interpretation	145 (87%)	46 (87%)	125 (95%)	52 (95%)
[Konfidenzintervalle]	[82%-92%]	[72%-94%]	[92%-99%]	[84%-99%]
..... richtig (eindeutige Lesart)	138,5 (83%)	44(83%)	122 (93%)	43,5 (79%)
..... richtig (mehrdeutige Lesart)	1	1	1	0
..... falsch	2,5	0	0	0,5
..... NIL	3	1	2	8 (15%)
... ohne Interpretation	22 (13%)	7 (13%)	6 (5%)	3 (5%)
..... Domänenmodell	–	–	–	1
..... Generisches Modell	1	5 (9%)	1	2
..... andere Domänen	1	–	–	–
..... Quantifikation	4	–	1	–
..... Abstrakta	4	–	–	–
..... Raum	–	–	2	–
..... Zeit	–	1	–	–
..... figurative Sprache	12 (7%)	1	2	–
..... Sonstiges	–	–	–	–
interrater reliability (n = 2)	$\kappa = 1$	$\kappa \approx 1$	$\kappa \approx 1$	$\kappa = 1$
Signifikanz				

Tabelle 6.4: Verstehensleistung bei Modal- und Auxiliarkonstruktionen

darstellt und damit kein physikalisches Objekt ist, nicht als Rollenfüller am entsprechenden Verbkonzept (OBTAIN für “erfaßt”) zugelassen. Als Interpretationsergebnis kommt dementsprechend *NIL* heraus.

Beim Rating der Ergebnisse, die die semantische Interpretation berechnet hat, ergaben sich sowohl beim Trainingsset als auch beim Testset eine hohe Übereinstimmung, was angesichts der klaren Abbildungsvorschrift allerdings wenig überraschend ist. Für das Trainingsset in der IT und das Testset in der Medizin ergab sich eine vollständige Übereinstimmung, im Testset der IT und dem Trainingsset der Medizin indessen eine Abweichung in jeweils nur einer Bewertung. Da ohnehin fast nur *richtig (eindeutig)* als Bewertung zugewiesen wurde, bewegt sich der Erwartungswert für das zufällige Raten sehr nahe bei 1. Aus diesem Grund ist eine κ -Statistik – wie bereits bei den Genitivattributen – nicht sehr aussagekräftig und der daran geknüpfte Signifikanztest nicht sinnvoll.

Qualifizierte Aussagen darüber, ob die Anzahl der Knoten in einem minimalen Subgraphen mit der Qualität der Interpretationsergebnisse korreliert, lassen sich aufgrund einer vergleichsweise sehr geringen Häufigkeit in den verwendeten Textsets leider nicht treffen. In der Medizin gibt es im Ganzen in beiden Sets nur einen einzigen Fall, in dem mehr als drei Knoten in einem minimalen interpretierbaren Subgraphen anzutreffen sind. Dieser konnte dann auch nicht analysiert werden, da das zur Interpretation unabdingliche syntaktische Subjekt fehlte. In der IT finden sich im Trainings- und Testset insgesamt 14 Okkurrenzen. Elf hiervon konnten interpretiert werden, wobei einmal *NIL* und zehnmal ein eindeutiges und richtiges Ergebnis herauskam. In einem Fall fehlte ein Konzeptidentifizier und in zwei Fällen das syntaktische Subjekt. Jedoch existieren aus früheren Arbeiten (Romacker et al., 1999) Evidenzen dafür, daß die Qualität der Interpretation unabhängig von der Anzahl der Knoten im Subgraph ist.

6.5.2 Evaluierung von Präpositionalphrasen

Präpositionalphrasen und deren disambiguierte Interpretation stellen eines der prominentesten Probleme beim Sprachverstehen dar. Allein die Häufigkeit von Präpositionen in einem Text sind ein eindeutiges Indiz hierfür. Gemessen an der Gesamtzahl der in den Trainings- und Testsets von Informationstechnologie und Medizin enthaltenen Wörter ist mehr als jedes zehnte Wort eine Präposition. Nimmt man hierzu noch für eine sehr defensive Abschätzung an, daß jede Präposition nur ein abhängiges Wort besitzt und von ein

weiteres Wort modifiziert, dann hängt mindestens ein Viertel des Textverständnisses bei den betrachteten Texten von der korrekten Interpretation der Präpositionalphrasen ab. Allerdings korreliert mit dem hochfrequenten Auftreten von Präpositionalkonstruktionen ein recht unterschiedlicher Gebrauch und damit eine unterschiedliche Funktion im Satz (Engel, 1988). Entsprechend lassen sich Klassen von Präpositionalphrasen bilden. Eine sehr große Teilmenge dieser Klassen läßt sich tatsächlich auch *semantisch* interpretieren, wenn man unter semantischer Interpretation versteht, daß eine Präposition eine semantische Beziehung zwischen den Referenzobjekten ihres semantischen Kopfs und ihres Modifikators anzeigt, die es anhand der im Weltwissen vorhandenen Objekts- und Eventdefinitionen zu berechnen gilt.

Da die Datengrundlage dieser Evaluation aus authentischen Texten besteht, muß – bevor mit der eigentlichen Evaluation der semantischen Interpretation begonnen werden kann – geklärt werden, welche Präpositionalphrasen überhaupt in die Kategorie semantisch interpretierbarer Präpositionalphrasen fallen. Das nachfolgende Kapitel 6.5.2.1 nimmt sich dieses Themas an, während die eigentliche Auswertung der Interpretationen im darauffolgenden Kapitel 6.5.2.2 erfolgt.

Abgesehen von der Beschränkung auf semantisch interpretierbare Präpositionalphrasen gibt es in dieser Klasse weitere Restriktionen dafür, ob sie in die Evaluation aufgenommen wurden. Eine Reihe von Phänomenen ist durch den Stand der Implementation in keinem evaluationsfähigen Maße abgedeckt. Dazu gehören Präpositionalphrasen, deren Modifikator ein Adverb (Bsp.: ab morgen) oder ein Adjektiv ist (Bsp.: ohne weiteres) sowie Adjektive, die als syntaktischer Kopf fungieren, sofern sie nicht Deverbalisierungen sind (Bsp.: stolz sein auf jemanden). Genaues Zahlenmaterial darüber, wieviele Präpositionalphrasen dadurch letztlich nicht interpretierbar sind, findet sich ebenfalls in Kapitel 6.5.2.2.

6.5.2.1 Klassifikation von Präpositionalphrasen

In einer groben Unterscheidung bezüglich der semantischen Interpretation lassen sich Präpositionalphrasen in vier Klassen aufteilen:

- Präpositionalphrasen ohne direkte Bedeutung, die syntaktisch gesehen koordinierende oder konjunktionsähnliche Funktionen wahrnehmen (z.B. auf der einen Seite, auf der anderen Seite).¹¹

¹¹Der Bedeutungsbegriff in unserem Kontext bezieht sich auf die Frage, ob einem Wort oder einem

- Präpositionalphrasen in mehr oder weniger *feststehenden Gefügen*, denen als Ganzes eine Bedeutung zuzuweisen ist. Diese sind im allgemeinen daran zu erkennen, daß es ein synonymes einzelnes Wort gibt (z.B. jemandem zur Seite stehen $\Leftrightarrow$ unterstützen; zum Teil $\Leftrightarrow$ teilweise). In der Systemphilosophie von SYNDIKATE erhält das Gefüge als Ganzes einen einzigen Konzeptidentifizier.
- Präpositionalphrasen, die als Ganzes selbst wieder *präpositionalen Charakter* haben (vgl. Engel (1988)). Von diesen Präpositionalphrasen hängt wiederum eine weitere Präpositionalphrase oder ein Genitivattribut ab. Durch die gesamte Präpositionalphrase werden zwei Inhaltswörter in eine durch die eingeschobene Präpositionalphrase als ganze vermittelte semantische Beziehung gerückt (z.B.: in Zusammenhang mit, auf der Ebene von/ auf der Ebene des etc.; diese würden als Ganzes auf eine Relation wie HAS-CIRCUMSTANCE abgebildet).
- Die letzte Klasse an Präpositionalphrasen umfaßt zahlenmäßig die meisten Okkurrenzen in den Texten: Eine Präposition verbindet syntaktisch zwei Inhaltswörter und schränkt die semantische Beziehung zwischen diesen durch die ihr innewohnende Bedeutung ein. Diese Präpositionalphrasen bilden einen minimalen interpretierbaren Subgraphen. In SYNDIKATE berechnet die semantische Interpretation die in diesem ausgedrückte Bedeutung. Die letzte Klasse teilt sich nochmals in zwei Kategorien auf: das Präpositionalobjekt und das Präpositionalattribut.

Aus den in der Aufzählung enthaltenen Beschreibungen wird ersichtlich, daß nur die letzte Klasse in den Bereich der semantischen Interpretation fällt, alle anderen sind bereits auf der Ebene des Lexikons zu behandeln, denn sie sind entweder überhaupt nicht instantiierbar (es existiert kein konzeptuelles Korrelat) oder sie sind als Mehrwortlexeme zu instantiieren (für das *gesamte* Präpositionalgefüge gibt es *eine* Menge an Konzeptidentifiern).

Um nun die in den der Evaluation zugrundeliegenden Texten enthaltenen interpretationsrelevanten Präpositionalphrasen zu identifizieren, wurde eine weitere empirische Studie vorgenommen. Fünf Studentinnen aus dem Grundstudium des Fachs Computerlinguistik, die keinerlei Kenntnis vom Textverstehenssystem SYNDIKATE besaßen, klassifizierten die Präpositionalphrasen der IT in den Texten nach einem vorgegebenen Klassifikationsche-

Gefüge aus mehreren Wörtern ein Menge von Konzeptidentifiern zugeordnet werden kann.

ma. In diesem Schema sind insgesamt fünf Kategorien enthalten, die im wesentlichen die vorangegangene Klassifikation widerspiegeln.¹² Bei der Vielfalt und Uneinheitlichkeit von Phänomenen, die an Präpositionen gebunden sind, stellen diese Kategorien eine starke Vereinfachung dar. Allerdings bringen sie genau in den Bereichen die nötige Trennschärfe mit sich, die zur Identifikation der zu verwendenden Interpretationsstrategie im Textverstehenssystem ausreicht. Damit konzentrieren sie sich auf die im gegebenen Rahmen notwendigen Verallgemeinerungen und nicht auf sprachliche Idiosynkrasien. Bei einem κ von 0.34091 (Signifikanzebene $\alpha = 0.01$ mit Signifikanzgrenze $z = 2.32$) ergab sich ein z von 16,6, so daß die Klassifizierer beim Rating signifikant übereinstimmten. Offensichtlich boten die definierten Unterscheidungskriterien ausreichenden Anhalt, um die Präpositionalphrasen mit signifikanter Übereinstimmung den Klassen zuzuordnen, d.h., die zur adäquaten semantischen Interpretation notwendige Aufteilung der Sprachdaten läßt sich empirisch absichern.

Die Kategorien eins bis drei im Klassifikationsschema beinhalten Präpositionalphrasen, die nicht interpretiert werden müssen, weil sie keine Inhaltswörter sind, oder die als Ganze auf eine einzelne Wortbedeutung abzubilden sind (erster bis dritter Punkt in der vorangehenden Auflistung; siehe insbesondere Anhang B). Für eine semantische Interpretation kommen allein die Kategorien vier und fünf in Betracht (vierter Punkt in der vorangehenden Auflistung), weshalb diese einer genaueren Analyse zu unterwerfen sind.

Im wesentlichen geben die Kategorien vier und fünf an, wie eine Präpositionalphrase zu interpretieren ist, denn in Abhängigkeit von der Kategorie kommt eine andere semantische Interpretationsfunktion zum Einsatz.¹³ Zunächst können Präpositionalphrasen Kasusobjekte einführen, die dann analog zu einem direkten oder indirekten Objekt zu interpretieren sind (Klasse 5). In Satz 18 bindet sich der PATIENT der Handlung (*“Mutter”*) mittels einer Präpositionalphrase (*“um seine Mutter”*) an das Verb. Entsprechend werden die statischen Constraints in diesem Fall auf das *direct-object* abgebildet. Dies wird im Lexikon auf der Ebene des Worteintrags kodiert, indem es eine ausgezeichnete Valenz für eine *um*-Präpositionalphrase am Verb *“kümmern”* gibt, die in der Restriktionsliste *direct-object* enthält.

¹²Eine genaue Auflistung und Beschreibung des Klassifikationsschemas findet sich im Appendix B. Die in den Medizin-Texten enthaltenen Präpositionalphrasen sind, da hierzu nicht verfügbare Experten sowohl in Medizin als auch in Linguistik erforderlich gewesen wären, von einem Mediziner und dem Autor zusammen nach dem Konsensprinzip klassifiziert worden.

¹³Diese Zusammenhänge waren den Klassifizierern natürlich unbekannt.

18. Hans kümmert sich um seine Mutter.

Im Gegensatz dazu verbieten die adverbial bzw. attributiv gebrauchten Präpositionalphrasen genau diese Form der Interpretation (Klasse 4). Daher finden sich in der Negativliste der entsprechenden Valenz *ppadverb* bzw. *ppattribute* die drei Elemente {*subject*, *diobject*, *indirobject*}, die über die statische Interpretation Lesarten wie AGENT oder PATIENT ausschließen. Die Entscheidung, welcher der beiden Fälle tatsächlich vorliegt, ist nicht immer einfach. Satz 19 illustriert die Problematik.

19. Der Rechner wird mit neuesten Technologien ausgestattet.

Die Präpositionalphrase “*mit neuesten Technologien*” hat eine ambige Interpretation. Bezeichnen die “*Technologien*” Hardwareteile, die in einen Computer eingebaut werden, dann wäre die Präpositionalphrase unter 5 zu klassifizieren. Meint man jedoch “*mit Hilfe modernster Technologien*”, dann ist die Phrase eindeutig der Klasse 4 zuzuordnen, da die Art und Weise des “*Ausstattens*” bezeichnet ist.

6.5.2.2 Ergebnisse der semantischen Interpretation von Präpositionalphrasen

Die Anzahl der Präpositionalphrasen, die vom Trainings- bzw. Testset ausgeschlossen wurden, weil sie keinen substantivischen Modifikator besitzen, belief sich im Trainingsset der IT auf 15 und im Trainingsset der Medizin auf 11. Im Testset ergaben sich als Werte 8 für die IT sowie 6 für die Medizin. Gemessen an der Gesamtzahl der Präpositionalphrasen dominieren also in beiden Domänen eindeutig Phrasen mit abhängigen Substantiven.¹⁴

Weiterhin nicht mit einbezogen in die empirische Untersuchung der semantischen Interpretation sind Präpositionalphrasen mit adjektivischem Kopf, es sei denn, diese adjektivischen Formen sind als Partizip von einem Verb abgeleitet.

20. Das von Lymphozyten *infiltrierte* Stroma.

21. Die in der Mucosa *anzutreffenden* Lymphozyten.

22. Die Läsion ist *verdächtig* auf eine Helicobacter-pylori assoziierte Gastritis.

¹⁴Diese wenigen Fälle erscheinen dann auch nicht in der Gesamtstatistik zu Präpositionen in Tabelle 6.6, die die Präpositionen und ihre absolute Häufigkeit in den Textsets auflistet.

Dementsprechend sind die Sätze 20 und 21 mit einem Partizip Perfekt Passiv bzw. einem Partizip Präsens Aktiv in der Untersuchung enthalten, während der Satz 22 nicht dazu gehört.

Nachdem nun eine ganze Reihe von Präpositionalkonstruktionen von der Betrachtung ausgeschlossen wurden, weil sie keine semantische Interpretation auslösen, führt dieses Kapitel die Ergebnisse der Interpretationen der verbliebenen Okkurrenzen von Präpositionalphrasen an.

Sehr interessant ist ein Blick auf die Häufigkeiten von Präpositionen in den Texten, die weiter unten ausführlich diskutiert werden. Da dem erhobenen Zahlenmaterial Texte mit sehr unterschiedlichem Zweck (Produktberichte und Befundberichte) und aus unterschiedlichen Domänen zugrundeliegen, sind die Beobachtungen im Großen und Ganzen verallgemeinerbar. Von den weit über hundert in Grammatiken erfaßten Präpositionen kommt offensichtlich nur ein kleinerer Teil in den Texten tatsächlich auch vor. In der Häufigkeit dominieren deutlich die in ihrer Bedeutung und Funktion ambigen Präpositionen *auf*, *für*, *in*, *mit*, *von* und *zu*.

Zieht man den sich aus der Verwendungshäufigkeit in den Texten ableitbaren Stellenwert von Präpositionalphrasen in Betracht, dann erscheint es folgerichtig, daß Interpretationsprozesse auf einer ganzen Reihe unterschiedlicher Ebenen aufsetzen müssen. Der in diesem Kapitel empirisch untersuchte Ansatz zur semantischen Interpretation stellt nur einen von diesen Interpretationsansätzen dar, der nicht alle Phänomene zu interpretieren in der Lage ist. Entsprechend negativ wirkt sich dies auf den Recall aus, der für Präpositionalphrasen deutlich gegenüber Genitivattributen bzw. Modal- und Auxiliärverben abfällt.

In der IT erreicht das Trainingsset (siehe Tabelle 6.5) noch einen Recall von 63% – und dies unter der Randbedingung, daß die systemseitig vorgegebenen Interpretationsmöglichkeiten optimal ausgenutzt werden. Nimmt man dann noch hinzu, daß die Precision ebenfalls nur den “niedrigen” Wert von 77% erreicht, zeigen sich deutlich die Grenzen des verwendeten Ansatzes zur semantischen Interpretation. Allerdings stellt sich in diesem Zusammenhang die bereits angedeutete Frage, ob nicht ein großer Teil der nicht erfaßten Okkurrenzen anderen Interpretationsschichten zuzurechnen ist. Gegenüber dem Trainingsset fällt das Testset in der IT mit einem Recall von 54% und einer Precision von 66% nochmals zurück.

Im Gegensatz zur IT fallen die Zahlen in der Medizin nicht ganz so schlecht aus. Im Trainingsset hält sich der Recall bei guten 77%, im Testset immerhin noch bei 74%. Das Testset

	Training-IT	Test-IT	Training-Med	Test-Med
Recall	63%	54%	77%	74%
Precision	77%	66%	79%	81%
# Vorkommen ...	611	254	562	278
... mit Interpretation	500 (82%)	206 (81%)	548 (98%)	253 (91%)
[Konfidenzintervalle]	[78%-84%]	[75%-85%]	[96%-99%]	[86%-93%]
..... richtig (eindeutige Lesart)	356 (58%)	119,5 (47%)	401,5 (71%)	167 (60%)
..... richtig (mehrdeutige Lesart)	30 (5%)	16,5 (6%)	32,5 (6%)	37,5 (13%)
..... falsch	42 (7%)	27 (11%)	43 (8%)	30,5 (11%)
..... NIL	72 (12%)	43 (17%)	71 (13%)	18 (6%)
... ohne Interpretation	111 (18%)	48 (19%)	14 (2%)	25 (9%)
..... Domänenmodell	11	15 (6%)	3	16 (6%)
..... Generisches Modell	52 (9%)	16 (6%)	5	6
..... andere Domänen	9	4	–	–
..... Quantifikation	4	–	5	1
..... Abstrakta	10	4		1
..... Raum	3	3	1	1
..... Zeit	2	–	–	–
..... figurative Sprache	19 (3%)	6	–	–
..... Sonstiges	1	–	–	
interrater reliability (n = 2)	$\kappa = 0,472$	$\kappa = 0,800$	$\kappa = 0,591$	$\kappa = 0,774$
Signifikanz	$z = 4,53$	$z = 6,31$	$z = 5,72$	$z = 6,41$

Tabelle 6.5: Verstehensleistung bei Präpositionalphrasen

weist sogar mit 81% eine allerdings unbedeutend bessere Precision verglichen mit dem Trainingsset auf, das auf immerhin 79% kommt. Prinzipiell dürften die besseren Ergebnisse in der Medizin auf den klaren und deutlich engeren thematischen Fokus zurückzuführen sein. Die Texte der IT decken ein deutlich weiteres inhaltliches Spektrum ab, was sich unter anderem dadurch belegen läßt, daß sich in den Textsets Artikel zu Telefonanlagen und Digitalkameras befinden, die außerhalb der eigentlichen computerbezogenen IT-Domäne liegen.

Ein Blick über alle vier Textsets zeigt, daß der Anteil an mehrdeutigen Interpretationen zwischen 5% und 13% liegt und damit die Restriktionslisten bei der semantischen Interpretation auch für Präpositionalphrasen das Anwachsen von interpretatorischen Mehrdeutigkeiten zu begrenzen in der Lage sind. Allerdings bringen die Constraints der Restriktionslisten auch ein Anwachsen der Interpretationen ohne Ergebnis (*NIL*) mit sich. Auffällig ist weiterhin, daß der schlechtere Recall in der IT auf fehlende und zum Teil wegen unscharfer Begriffe nicht realisierbare Konzeptdefinitionen im UPPER-MODEL zurückzuführen ist (9% im Trainingsset bzw. 6% im Testset); in der Medizin liegt der Anteil dieser Begriffe im Trainingsset sehr niedrig, erreicht aber im Testset das Niveau der IT. Daneben scheint ausschließlich in der IT figurative Sprache mit 19 Okkurrenzen (3%) noch eine kleinere Rolle zu spielen. Da sich bei Präpositionalphrasen eine statistisch verwertbare Verteilung der Phänomene auf die Bewertungskategorien *falsch* und *richtig* mit deutlichen Abweichungen zwischen den Ratern findet, ist die κ -Statistik in Tabelle 6.5 angegeben. Die Inter-Rater-Reliability übersteigt in allen vier Sets bei unterschiedlichem Kappa-Wert das Signifikanzniveau von $z = 2.32$ für $\alpha = 0.01$ (Siegel & Castellan, 1988).

Tabelle 6.6 enthält sämtliche in den Textsets enthaltenen Präpositionen, von denen ein Substantiv abhängt, sowie deren Häufigkeit. Dabei weisen die Zellen der Tabelle sowohl die absolute Zahl als auch die Zahl der in der Evaluation der semantischen Interpretation tatsächlich vorkommenden Präpositionen (in Klammern) aus. Der zahlenmäßige Unterschied ergibt sich aus der Anzahl der tatsächlichen Okkurrenzen abzüglich derjenigen Präpositionalphrasen, deren semantische Interpretation nicht erforderlich ist, da es sich um ein im Lexikon zu lösendes Abbildungsproblem von mehreren Wortoberflächen auf *eine* Menge an Konzeptidentifiern handelt und damit keine Instantiierung semantischer Interpretationsfunktionen nötig ist. Nur bei den ebenfalls von der Empirie ausgeschlossenen adjektivischen Köpfen, die sich nicht von einem Verb ableiten, müßte eigentlich eine semantische Interpretation stattfinden. Trotz dieser Einschränkungen sind von den insge-

samt 1978 Präpositionalphrasen immerhin 1705 (86%) in der Datengrundlage der Empirie anzutreffen. Ohne in extenso auf die einzelnen Daten einzugehen, sei als prominentes Beispiel die Präposition „zu“ im Medizin-Trainingsset herausgegriffen. Von den insgesamt 33 Okkurrenzen konnten gerade sieben Mal interpretiert werden. Dies ist indessen darauf zurückzuführen, daß über 20 Mal die Wendung „zum Teil“ auftritt, die durch das Adverb *teilweise* paraphrasierbar und damit im Lexikon zu handhaben ist. Zusätzlich sei angemerkt, daß aufgrund der Negation, die in der Präposition „ohne“ eingeschlossen ist, eine semantische Interpretation von Präpositionalphrasen mit „ohne“ nicht möglich ist. Eine Verarbeitung von Negation unterbleibt beim augenblicklichen Stand der Implementation in SYNDIKATE.

Betrachtet man insgesamt die Verteilung der Präpositionen, dann fällt auf, daß insgesamt überhaupt nur ein kleiner Teil aller existierenden Präpositionen verwendet wird (ca. 50). In einschlägigen Grammatiken (Engel, 1988; Drosdowski, 1984) finden sich mehrere Hundert. Auch die tatsächlich verwendeten Präpositionen haben eine höchst ungleichmäßige Distribution. Die Präpositionen *für*, *in*, *mit* und *von* stellen mit insgesamt über 1.300 Okkurrenzen bereits 66% aller Vorkommen. Eine sehr große Anzahl an Präpositionen bewegt sich in einem marginalen Bereich. Fast die Hälfte aller Präpositionen tritt nicht mehr als zehnmal auf. Eine Frage, die sich an die Distribution knüpft, ist die, in welcher Form die jeweiligen Präpositionen zur Bedeutung der Phrasen beitragen. Aufgrund der enormen Datenmenge kann dieser Fragestellung hier nur an prototypischen Beispielen nachgegangen werden. Zunächst folgt eine qualitative Beschreibung typischer Phänomene, ohne diese zu quantifizieren. Anschließend liefert eine quantitativ-qualitative Beschreibung konkretes Datenmaterial.

6.5.2.3 Qualitative Analyse der empirischen Daten zur Präpositionalphrasen

In den empirischen Daten zur semantischen Interpretation finden sich eine Reihe von Evidenzen, die prinzipielle Entscheidungen zur Systemarchitektur legitimieren. In diesen Bereich fällt das Matchen von Instanzen auf der assertionalen Ebene, bei dem mittels assertionaler Axiome, die aus der Interpretation gewonnen werden, eine Instanz auf einen spezielleren Konzepttyp abgebildet wird. Typischerweise korreliert dieser Sachverhalt mit Textualitätsphänomenen. Da der dynamische Teil der semantischen Interpretation auf dem Relationsinventar des Ausgangskonzepts aufsetzt, kommt dem Matchen von Instanzen auf speziellere Konzepte eine wichtige Rolle beim Interpretationsprozeß zu.

Präposition	IT-Training	Med-Training	IT-Test	Med-Test	Summe
ab	5 (5)	2 (-)	4 (3)	- (-)	11 (8)
als	12 (-)	- (-)	4 (-)	- (-)	16 (-)
an	21 (19)	16 (10)	7 (7)	8 (5)	52 (41)
anlässlich	1 (-)	-	-	-	1 (-)
auf	54 (41)	12 (9)	14 (14)	6 (2)	84 (66)
aufgrund	- (-)	1 (1)	1 (1)	- (-)	2 (2)
aus	7 (7)	30 (28)	3 (3)	21 (21)	61 (56)
bei	28 (25)	13 (11)	12 (12)	6 (5)	59 (50)
bis	13 (12)	21 (-)	7 (7)	11 (-)	52 (19)
dank	1 (1)	- (-)	- (-)	- (-)	1 (1)
durch	6 (6)	10 (8)	1 (1)	6 (6)	23 (21)
einschließlich	1 (1)	- (-)	- (-)	- (-)	1 (1)
entsprechend	- (-)	1 (1)	1 (-)	- (-)	2 (1)
für	90 (87)	50 (47)	25 (24)	26 (26)	191 (184)
gegen	1 (-)	- (-)	2 (2)	- (-)	3 (2)
gemäß	1 (1)	- (-)	- (-)	- (-)	1 (1)
in	85 (81)	159 (134)	48 (43)	75 (69)	367 (327)
inklusive	3 (3)	- (-)	1 (1)	- (-)	4 (4)
innerhalb	- (-)	1 (1)	- (-)	- (-)	1 (1)
jenseits	1 (1)	- (-)	- (-)	- (-)	1 (1)
laut	4 (4)	1 (1)	2 (2)	- (-)	7 (7)
mit	140 (135)	170 (170)	71 (59)	82 (81)	463 (445)
mittels	- (-)	5 (5)	- (-)	- (-)	5 (5)
nach	6 (2)	11 (9)	3 (3)	4 (4)	24 (18)
neben	- (-)	1 (1)	- (-)	2 (2)	3 (3)
oberhalb	- (-)	- (-)	- (-)	1 (1)	1 (1)
ohne	3 (-)	10 (-)	6 (-)	8 (-)	27 (-)
per	1 (1)	- (-)	2 (2)	- (-)	3 (3)
pro	3 (-)	- (-)	1 (-)	- (-)	4 (-)
seit	1 (1)	- (-)	1 (1)	- (-)	2 (2)
statt	1 (1)	- (-)	1 (1)	- (-)	2 (2)
über	17 (17)	7 (7)	9 (9)	4 (4)	37 (37)
um	12 (11)	9 (9)	5 (5)	- (-)	26 (25)
unter	9 (8)	7 (-)	4 (4)	2 (1)	22 (13)
via	- (-)	- (-)	2 (2)	- (-)	2 (2)
von	98 (97)	105 (103)	39 (34)	44 (42)	286 (276)
vor	2 (1)	- (-)	2 (2)	- (-)	4 (3)
wegen	1 (1)	- (-)	- (-)	1 (1)	2 (2)
zu	47 (38)	33 (7)	9 (9)	13 (8)	102 (62)
zwischen	4 (4)	- (-)	6 (3)	- (-)	10 (7)
Summe	679 (611)	675 (562)	291 (254)	333 (278)	1978 (1705)

Tabelle 6.6: Verteilung der Präpositionen in den Trainings- bzw. Testsets

Im Satz 23 triggert die semantische Interpretation eine Anfrage, wie das konzeptuelle Korrelat von *“Schleimhaut”* (MUCOSA) mit der Typqualifikation *“Antrumtyp”* relationierbar ist. In der medizinischen Ontologie hat jedoch erst das Subkonzept DIGESTIVE-MUCOSA von MUCOSA eine entsprechende Rolle (HAS-MUCOSA-TYPE). Die semantische Interpretation liefert folglich kein richtiges Ergebnis, sondern gibt *NIL* zurück.

23. Die Schleimhaut vom Antrumtyp.

Die Auflösung der Referenzen im hier nicht dargestellten textuellen Kontext – im Vor-text wird ein *“Magen”* eingeführt – ermöglicht die Berechnung der richtigen Relationierung dadurch, daß *“die Schleimhaut”* als anaphorischer Ausdruck interpretiert wird. Durch die Anbindung des Antezedenten *“Magen”* als textuelle Ellipse an *“Schleimhaut”* mittels der Relation HAS-ANATOMICAL-LOCATION kann der Recognizer den deduktiven Schluß ableiten, daß die Instanz für *“Schleimhaut”* in der Textwissenbasis dem spezielleren Typ GASTRIC-MUCOSA zugehört, der wiederum von DIGESTIVE-MUCOSA subsumiert wird. Nur durch diesen, vom Wissensrepräsentationssystem geleisteten Ableitungsschritt kann die richtige Relation HAS-MUCOSA-TYPE berechnet werden.

Die große Vielfalt bei der Verwendung von Präpositionalphrasen bringt indessen mit sich, daß selbst die Resolution textueller Phänomene nicht notwendig eine richtige semantische Interpretation ermöglicht. Der Satz 24 enthält ebenfalls eine textuelle Ellipse, deren Auflösung den Diskursreferenten *“Drucker”* in den Satzkontext einführt: *“Der Preis des Druckers inklusive Druckerlabel beträgt 1.200 DM”*. Nach rein semantisch-ontologischen Kriterien müßte die Präpositionalphrase *“inklusive Druckerlabel”* an den nicht verfügbaren Antezedenten angebunden werden. Die syntaktisch erreichbaren Köpfe *“Preis”* und *“beträgt”* führen zu keiner richtigen semantischen Interpretation. Im Resultat bleiben Sätze dieser Art uninterpretierbar.

24. Der Preis inklusive Druckerlabel beträgt 1.200 DM.

Als weiteres interpretationskritisches textuelles Phänomen spielt die nominale Anapher eine Rolle, um ein richtiges Interpretationsergebnis zu berechnen. Der Satz 25 gibt ein Beispiel aus dem IT-Trainingsset dafür, daß die Auflösung der nominalen Anapher Voraussetzung einer erfolgreichen semantischen Interpretation ist.

25. Produkte mit SGRAM.

Der textuelle Kontext handelt allgemein von Grafikkarten (GRAPHIC-CARD) und vergleicht diese miteinander. Damit stellt *“Produkte”* einen anaphorischen Ausdruck – übrigens ohne Markierung als definite Nominalphrase – für eine Menge von Antezedenten dar, die im Vortext beschrieben sind. Eine semantische Interpretation mit dem Konzepttyp PRODUCT muß scheitern, da dieses Konzept nur sehr allgemeine Eigenschaften wie HAS-PRICE oder PRODUCED-BY definiert. SGRAM-MODULE als Subkonzept von Arbeitsspeicher ist zu speziell, um mit PRODUCT relationierbar zu sein. Allein mittels der referentiellen Auflösung von PRODUCT zu GRAPHIC-CARD läßt sich ein gültiger Relationspfad ermitteln. Dies unterstreicht nochmal eindringlich den Stellenwert der Anaphernresolution gerade für das grundsätzlich nur lokal agierende Verfahren zur semantischen Interpretation, da dieses zunächst ausschließlich auf den unmittelbar verfügbaren Typinformationen der beteiligten Instanzen aufsetzt.

Die Probleme, die mit der Interpretation von Präpositionalphrasen verbunden sind, erweisen sich indessen in einer zahlenmäßig nicht unerheblichen Anzahl von Fällen als nahezu unlösbar. Offensichtlich bieten bestimmte präpositionale Konstruktionen die Möglichkeit, Sachverhalte extrem ökonomisch mit einem großen Grad an Implizitheit auszudrücken (Schauer, 2000), der formal kaum rekonstruierbar ist. In Satz 26 beinhaltet die Präpositionalphrase eine komplexe Aktion, die man aus dem Kontext im Text heraus folgendermaßen paraphrasieren könnte: *Kein Käufer verläßt das Geschäft, ohne eine Grafikkarte gekauft zu haben, deren Speicherbausteine mindestens 8 MByte Kapazität haben.*

26. Unter 8 MByte wird kein Käufer heim geschickt.

Isoliert betrachtet macht dieser Satz eigentlich keinen Sinn und verläßt möglicherweise schon den Rahmen der Akzeptabilität. In seinen textuellen Kontext eingebettet jedoch, ist er problemlos verständlich und damit akzeptabel.

Die Interpretationsergebnisse bei Präpositionalphrasen fallen auch deshalb gegenüber den anderen untersuchten Sprachphänomenen zurück, weil sich ein Teil der generierten Anfragen gar nicht auf eine Menge an Relationspfaden, die sich aus dem Konzeptgraphen errechnet, abbilden läßt. Schauer & Hahn (2000) weisen in ihrer Untersuchung zur Funktion von Cue-Phrases in Texten nach, daß ca. 10% der Präpositionalphrasen das Vorhandensein einer Kohärenzrelation anzeigen. Im allgemeinen stehen derartige Kohärenzrelationen wie EXPLANATION oder MOTIVATION aus der Rhetorical Structure Theory (vgl. Mann & Thompson (1988)) in keiner Verbindung zu den im Weltwissen kodierten konzeptuellen

Relationen; demzufolge berühren Kohärenzrelationen eine ganz andere Strukturierungsebene innerhalb eines Texts. Dennoch fällt es nicht immer leicht, eine klare Trennungslinie zu ziehen. Konkretere Kohärenzrelationen, die nicht nur Einheiten eines Texts strukturieren, sondern auch Entitäten einer Domäne zu relationieren geeignet sind, überschreiten die Grenze zu reinen Relationen des Weltwissens bereits. Die Kohärenzrelation PURPOSE beispielsweise könnte durchaus in Definitionen von Verb-Konzepten oder Objekten, vorzugsweise Artefakten, Eingang finden. Auffällig bei derartigen Relationen ist jedoch deren Allgemeingültigkeit: Welche Typbeschränkungen sind für den “Zweck” einer Handlung oder eines Objekts bezüglich möglicher Füller überhaupt festlegbar?

Die in SYNDIKATE vorgegebene Philosophie bei der Wissensmodellierung legt fest, daß Relationen in Konzeptdefinitionen mit sehr laxen Typenconstraints zu vermeiden sind, da die semantische Interpretation bei der Pfadsuche dann übergeneralisiert und eine größere Menge ambiger Ergebnisse entstehen. Aus diesem Grund finden sich im Weltwissen keine der in der RST definierten Relationen. In letzter Konsequenz bleibt die präzise Abgrenzung des epistemischen Unterschieds zwischen Kohärenz- und Weltwissensrelationen ein offenes Problem, wenn der unscharfe Überschneidungsbereich, der durch Relationen wie MEANS, PURPOSE oder CAUSE gegeben ist, einer genaueren Analyse unterworfen wird.

Betrachtet man in diesem Zusammenhang das Inventar an Präpositionen, dann gibt es idealtypisch unterschieden drei Klassen, wenn man die Unterscheidung in Weltwissens- oder Kohärenzrelation als Klassifikationskriterium zugrunde legt. Zunächst existieren Präpositionen, die eindeutig und ausschließlich die Präsenz einer Kohärenzrelation nahelegen. Die Präposition “*aufgrund*” führt beispielsweise eine EXPLANATION ein, während die Präposition “*anlässlich*” auf das Vorliegen einer CIRCUMSTANCE-Relation schließen läßt. Im Gegensatz dazu gibt es Präpositionen, die exklusive Weltwissensrelationen anzeigen. “*Seit*” gibt zum Beispiel ausschließlich eine temporale Relation zu einem Event an (START-TIME-INTERVAL). Die letzte Klasse setzt sich aus den mehrdeutigen Präpositionen zusammen, die sowohl eine Weltwissensrelation als auch eine Kohärenzrelation einführen können. Es sind jedoch gerade die hochfrequenten Präpositionen wie “*mit*”, “*für*” oder “*bei*”, die in diese Klasse fallen und damit auf der Ebene ihrer Funktion im Text zunächst ambig bleiben.

Deshalb ist in manchen Fällen die Entscheidung, ob eine Kohärenzrelation vorliegt oder eine Weltwissensrelation, nicht leicht zu treffen. Der Satz 27 enthält eine durch die Präposition “*für*” gegebene Präpositionalphrase, die gemäß dem Klassifikationsschema für Präpositionalphrasen als Präpositionalobjekt einzustufen ist. Eine semantische Interpre-

tation erfolgt entsprechend durch Abbildung auf die ACTION-PATIENT Relation am Verbkonzept für *“sorgen”*. Eine an Kohärenzrelationen orientierte Betrachtungsweise könnte beim gleichen Beispiel zu dem Ergebnis kommen, daß eine PURPOSE-Relation vorliegt. Insbesondere Verben, in denen eine telische Bedeutung eingeschlossen ist, erschweren die Beurteilung, welche Interpretationsebene greift, wobei durchaus parallele Interpretationsprozesse denkbar sind.

27. Ports sorgen für Flexibilität beim Uplink.

Ähnlich schwierig zu interpretieren ist die Präposition *“bei”*, an der sich ebenfalls die Koexistenz von Weltwissens- und Kohärenzrelationen festmachen läßt.

28. Der Scanner liest bei einer Farbtiefe von 24 Bit Vorlagen mit 300 x 600 dpi ein.

Die Präpositionalphrase *“bei Farbtiefe”* führt ganz klar eine CONDITION-Kohärenzrelation ein, d.h., wenn die Farbtiefe 24 Bit beträgt, dann wird die Auflösung von 300 x 600 dpi erreicht.¹⁵ Andererseits bezeichnet Farbtiefe auch eine, allerdings variierbare Eigenschaft des Scanners im Satz. Jedoch beziehen sich Weltwissens- bzw. Kohärenzrelationen auf je unterschiedliche Ebenen im Kommunikationsprozeß, so daß die potentielle Präsenz beider zunächst keinen Widerspruch darstellt.

Insgesamt betrachtet sind es gerade die Präpositionen, deren Bedeutung und Funktion im Äußerungskontext nur ungenau zu fassen ist, welche der semantischen Interpretation von Präpositionalphrasen erhebliche Schwierigkeiten bereiten und Recall sowie Precision verschlechtern.

Vergleicht man abschließend nochmals die Interpretationsergebnisse aus IT und Medizin, dann läßt sich der bessere Recall und die höhere Precision in der Medizin auch darauf zurückführen, daß die Texte – entsprechend dem kompakten Schreibstil der Mediziner – deutlich weniger durch Präpositionen angezeigte Kohärenzrelationen aufweisen. D.h., Präpositionalphrasen sind nahezu immer auf der Grundlage des Domänenmodells interpretierbar, und genau auf dieser Ebene setzt die eigentliche semantische Interpretation an.

¹⁵Die Kohärenzrelation CONDITION kann als eine semantisch *“schwache”* Relation zur Darstellung der Implikation angesehen werden.

6.5.2.4 Qualitativ-quantitative Analyse der empirischen Daten zur Präpositionalphrasen

Die qualitativ-quantitative Analyse der empirischen Daten setzt auf zwei Ebenen an. Zunächst kommen schlaglichtartig einige prominente Phänomene zweier Präpositionen zur Darstellung, ohne daß hierbei das Datenmaterial vollständig ausgewertet wurde. Aus der Menge der hochfrequenten Präpositionen seien stellvertretend *“aus”* und *“in”* herausgegriffen. Zum anderen sind die gewonnenen Daten für die beiden ebenfalls hochfrequenten Präpositionen *“mit”* und *“von”* einer umfassenden Analyse unterworfen worden. In diesen Betrachtungen stehen domänenspezifische Besonderheiten und Unterschiede zwischen den Domänen im Vordergrund. Obwohl mit insgesamt 1705 analysierten Präpositionalphrasen eine sehr große Anzahl verarbeitet werden konnte, sind wegen des seltenen Auftretens der meisten Präpositionen Aussagen über diese nicht zu treffen. Trotzdem gehen die folgenden Erörterungen von den tatsächlichen Daten aus und versuchen diese anhand der interpretatorischen Merkmale zu klassifizieren. Daß dabei notwendige Vereinfachungen zu treffen sind, resultiert aus der Tatsache, daß das Thema der semantischen Interpretation von Präpositionalphrasen sehr vielschichtig ist. Daher gibt es viele Monographien, die sich auf der deskriptiven Ebene mit nur einer einzigen Präpositionen oder mit Teilaspekten der Semantik von Präpositionen auseinandersetzen (zum Beispiel Kaufmann (1995), Sekiguchi (1994), Rauh (1991)).

Bezüglich der Präposition *“aus”* enthalten die Daten in der Medizin interessante Zusammenhänge. Bei den 28 im Trainingsset enthaltenen und der semantischen Interpretation zugeführten Präpositionalphrasen mit *“aus”* herrschen zwei Sprachmuster vor. Typischerweise bezeichnen die Präpositionalphrasen entweder *“stammen aus”* (Bezug zu einer anatomischen Region oder Struktur) mit 6 Okkurrenzen oder *“bestehen aus”* (meronymische Beziehung) mit 17 Okkurrenzen. Leider gibt es in diesen Fügungen sehr häufig Verbausslassungen, insgesamt 14 der 23 Fälle ließen das Verb aus. Dies kann als Indiz dafür gewertet werden, daß die Verben die in den Restriktionslisten der Präpositionen kodierten positiven Constraints SOURCE bzw. HAS-PART lediglich explizit machen. In diesem Zusammenhang gewinnt die konzeptuelle Interpretation an Bedeutung, da sie im Fall der Verbpräsenz eine Relationierung berechnet, die im Omissionsfall unmittelbar aus der Interpretation der Präpositionalphrase resultiert.

Im Vergleich der Medizin-Domäne mit der IT-Domäne bezüglich der Interpretationen

der Präposition *“in”* weist die Medizin eine erstaunliche Beschränkung auf. Die Bedeutungsrestriktionen für Präpositionalphrasen mit *“in”* sind im allgemeinen relativ weit gefaßt, neben dem lokativen Gebrauch führt *“in”* temporale Angaben über Zeitpunkte und Zeiträume ein, sie beschreiben modale Bezüge und vieles mehr. Die in der Informationstechnologie in den Textsets anzutreffenden *“in”*-Präpositionalphrasen bieten auch das gesamte Spektrum möglicher Interpretationen. Die Medizin indessen beschränkt sich in den Histologieberichten eindeutig auf die lokale Lesart. Von den 134 im Trainingsset der Medizin anzutreffenden Okkurrenzen zeigen insgesamt 130 eine lokale Beziehung an. Dies ist ein eindeutiges Indiz für den Fachsprachen Charakter der Medizin, der es ermöglicht, Textverstehenssysteme unter Beschränkung der Allgemeinheit an diese Domäne anzupassen.

Die Präposition *“mit”* ist mit 463 Okkurrenzen die häufigste in allen vier Textsets, und zwar sowohl in der Gesamthäufigkeit, als auch in jedem der einzelnen Textsets. Entsprechend vielfältig sind die konzeptuellen Beziehungen, die durch *“mit”* angezeigt werden und von der semantischen Interpretation aufzulösen sind. In den nachfolgenden Sätzen sind jeweils aus der IT und Medizin prototypische Beispiele gegeben, sofern sich Evidenzen in den Texten hierfür finden lassen. Hinter den Beispielen findet sich in Klammern die Angabe zur Häufigkeit in den Texten, wodurch die Relevanz eines Phänomens für die Interpretation abzuschätzen ist.

Die Sätze 29a und 29b greifen jeweils eine Phrase auf, in der eine über *“mit”* kodierte partitive Beziehung vorliegt. Ein Display ist genauso Teil eines Notebooks wie Zotten als Teil einer Schleimhaut anzusehen sind. Darüber hinaus binden Phrasen dieses Typs nicht nur Teile an ein Ganzes, sondern auch weitere Eigenschaften an den syntaktischen Kopf der Präpositionalphrase.

29. a. Das Pacomp-Notebook mit 10-Zoll-Farbdisplay. (57/135)
 b. Erfäßt ist eine Schleimhaut mit schlanken Zotten. (76/170)

In diesen Bereich fallen etwa Präpositionalphrasen des Typs *“ein 19-Zoll-Display mit einer Standardauflösung von 1024 x 768 (IT) bzw. “eine Magenschleimhaut mit regelrechter Architektur” (Medizin)*. In beiden Domänen treten *“mit”*-Präpositionalphrasen dieses Typs am häufigsten auf. Die Restriktionsliste der Präposition kodiert diesen Zusammenhang durch die Relationen HAS-PART und HAS-ATTRIBUTE.¹⁶

¹⁶Anhand dieses Beispiels wird deutlich, wie die Restriktionslisten für Präpositionen gewonnen wurden;

Ein in der Medizin hochfrequentes Sprachmuster findet sich in Satz 30. Obwohl beinahe die Hälfte aller *“mit”*-Präpositionalphrasen durch dieses Muster abgedeckt werden, findet sich in der IT erstaunlicherweise kein Pendant. Typischerweise assoziieren Befundberichte einen Befund (hier die Antrumgastritis) mit weiteren diagnostischen Beobachtungen (hier die Metaplasie), die übrigens nicht unbedingt kausal zusammenhängen müssen.

30. Eine Antrumgastritis mit intestinaler Metaplasie. (75/170)

In einer ähnlichen Konstruktion kombiniert ein anatomisches Konzept mit einem Befund: *“eine Antrumschleimhaut mit auffälliger lymphozytärer Infiltration”*. Interessanterweise bezeichnet die *“Infiltration”* einen pathologischen Prozeß, dessen PATIENT durch den syntaktischen Kopf angegeben ist. Um diesen medizintypischen Interpretationen Rechnung zu tragen, befinden sich in der Restriktionsliste von *“mit”* die Relationen HAS-CO-PHENOMENON und PATIENT-OF.

Die Sätze 31a und 31b enthalten Beispiele, bei denen die Präpositionalphrase vom Verb abhängt und zugleich als Präpositionalobjekt zu interpretieren ist. D.h., der von *“mit”* abhängige Modifikator füllt die Argumentstelle eines Kasusobjekts.

Dementsprechend erfolgt eine Abbildung auf PATIENT oder CO-PATIENT am dazugehörigen Verbkonzept. Kodiert sind die nötigen Constraints im Lexikon am Verb selbst, das für eine ausgezeichnete *ppobject*-Valenz eine Interpretation der Präpositionalphrase als *direct-object* bzw. *indirect-object* erfordert.

31. a. Toshiba wartet zur CeBit mit einem neuen Subnotebook auf. (32/135)

b. Die Kapillaren sind prall mit Erythrozyten angefüllt. (10/170)

Die INSTRUMENT-Lesart von *“mit”* kommt in der IT auf immerhin 16 Fälle, in der Medizin bleibt sie indessen ohne Okkurrenz. In Satz 32 findet sich ein Beispiel einer entsprechenden Präpositionalphrase, in dem die Firma *“Sony”* ein neues *“Subnotebook”* als INSTRUMENT verwendet, um Aufsehen zu erregen. Dementsprechend beinhaltet die Restriktionsliste von *“mit”* die Relation INSTRUMENT.

32. Mit einem Subnotebook sorgte Sony zur CeBit für Aufsehen. (16/135)

die in den Sprachdaten enthaltenen Muster liefern die empirischen Kriterien, anhand derer die Abstraktionen in der Relationshierarchie ableitbar sind. Das hierarchisch gesehen höchste Element eines Relationsbaums kommt dann in die Restriktionsliste.

Die nachfolgenden beiden Beispielsätze 33 und 34 weisen eine gewisse Ähnlichkeit auf. Darüber hinaus gibt es Evidenzen für Sätze dieses Typs nur in der IT, nicht aber in der Medizin.

33. Mit einer mageren Akkulaufzeit von 2,7 h eignet sich der Vaio nicht für den mobilen Betrieb. (16/135)

Auf den ersten Blick betrachtet könnten die in den beiden Sätzen enthaltenen Präpositionalphrasen als INSTRUMENT interpretiert werden. Jedoch weisen sie gegenüber der INSTRUMENT-Interpretation in Satz 32 einen klaren und in der semantischen Interpretation möglicherweise nicht unmittelbar zu berücksichtigenden Unterschied auf. In Satz 33 und Satz 34 besteht ein ontologischer Bezug zwischen dem Modifikator in der Präpositionalphrase und dem Subjekt bzw. dem direkten Objekt des Matrixsatzes. Die *“Akkulaufzeit”* referenziert klar eine Eigenschaft des Notebooks *“Vaio”*, wohingegen das Produkt von Texas Instruments, der *“PCI1221”*, eindeutig ein Subkonzept des durch die sprachliche Oberfläche *“PCI-to-Cardbus-Controller”* markierten Konzepts darstellt.¹⁷

34. Texas Instruments stellt mit dem PCI1221 einen PCI-to-Cardbus-Controller vor. (6/135)

Sätze dieser Art sind schwer zu interpretieren. Dennoch reicht die reine Interpretation als INSTRUMENT vielleicht für die semantische Interpretation aus. Weitergehende Inferenzen fallen in die Zuständigkeit der konzeptuellen Interpretation, die die Aufgabe hat, Constraints (hier: Relationspfad bzw. Subsumptionbeziehung) zwischen Rollenfüllern zu prüfen. Freilich gibt es auf der Ebene der konzeptuellen Interpretation bis jetzt keine auf dieses Phänomen zugeschnittene Interpretationsregeln. Bedauerlicherweise bringt dieses Manko mit sich, daß inkohäsive und inkohärente Textwissensbasen entstehen, da die impliziten Bezüge der Referenz (*“PCI1221”* und *“PCI-to-Cardbus-Controller”* sind referenzidentisch) und der Relationierbarkeit (*“Akkulaufzeit”* ist Eigenschaft von *“Vaio”*) nicht formal rekonstruiert werden. Jedoch beschränkt sich dieser Zusammenhang auf *“mit”*-Präpositionalphrasen, so daß das Problem – gegeben die Häufigkeit von 22 Okkurrenzen – vernachlässigbar ist.

¹⁷Schauer & Hahn (2000) interpretieren die *“mit”*-Präpositionalphrase als Kohärenzrelation (EXPLANATION). Damit deutet sich an, daß die semantisch reichhaltigeren Bezüge zwischen den Füllern am Verb *“eignet”* Evidenzen für eine Kohärenzrelation anzeigen können.

Im nächsten Beispielsatz 35 findet sich eine MANNER-Relation, die umschreibt, in welcher Art und Weise die *„Hersteller“* arbeiten.

35. Mit Hochdruck arbeiten die Hersteller an besserer Performance. (4/135)

Wegen ihres seltenen Auftretens und fehlender ontologischer Spezifikationen verarbeitet die semantische Interpretation MANNER-Präpositionalphrasen augenblicklich noch nicht. Dies begründet sich insbesondere durch die Nähe dieses Sprachmusters zu Adverbien und Adjektiven (*„intensiv“*, *„nachdrücklich“*), die generell von dieser Arbeit ausgeschlossen blieben. Darüber hinaus besteht ein Bezug zu RST, da MANNER dort als Kohärenzrelation verwendet wird (Mann & Thompson, 1988).

Die beiden letzten Beispielsätze zur Präposition *„mit“* enthalten Fälle, die schwer oder gar nicht interpretierbar sind, da sie entweder Sachverhalte extrem komprimieren, oder möglicherweise sogar semantisch inakzeptabel sind. Der Satz 36a legt in einer naiven Interpretation nahe, daß der glückliche Kunde vom Verkäufer mit 300 DM beschenkt wurde. Gemeint ist natürlich, daß der Kunde das Geschäft verläßt und insgesamt 300 DM ausgegeben hat – das genaue Gegenteil also. Den zur Ableitung dieses Zusammenhangs notwendigen Schlußfolgerungen ist indessen Wissen unterlegt, das mit den Konzeptbeschreibungen in SYNDIKATE nicht mehr erfaßbar ist, da es sich um detailliertes Regelwissen handelt. Folglich sind Sätze dieser Art nicht interpretierbar.

36. a. Der Käufer kommt mit 300 DM aus dem Geschäft. (4/135)
b. Biopsie mit einer unregelmäßig aufgebauten Schleimhaut. (9/170)

Der Beispielsatz aus der Medizin stellt die semantische Interpretation deshalb vor Probleme, weil eine *„Biopsie“* normalerweise ein Teil einer anatomischen Struktur – wie zum Beispiel einer *„Schleimhaut“* – *„erfaßt“*. Damit ist die Biopsie ein anatomisches Fragment der betroffenen anatomischen Struktur(en). Die angeführte Präpositionalphrase stellt diesen Zusammenhang auf den Kopf und suggeriert, die Schleimhaut sei Teil der Biopsie. Aus diesem Grund gibt die semantische Interpretation kein gültiges Ergebnis. Abschließend stellt sich die hier nicht klar zu beantwortende Frage, ob ein derartiger Satz überhaupt akzeptabel ist.

Insgesamt betrachtet kommen in der Medizin weniger Grundtypen bei der Verwendung von *„mit“*-Präpositionalphrasen vor. In diesem Zusammenhang bestätigen die empirischen

Daten, wie schon bei der Präposition *“in”*, daß die Medizin als Fachsprache eine restriktierte Menge von semantisch-konzeptuellen Verbindungen verwendet, was die Spezifikation an die Medizin adaptierter semantischer Grammatiken erleichtert bzw. überhaupt erst ermöglicht.

Als zweite Präposition wird nun *“von”* einer umfassenden Analyse unterzogen. Präpositionalphrasen mit *“von”* führen offensichtlich sowohl in der IT als auch in der Medizin sehr spezifische Sprachmuster ein, die jedoch aufgrund ihrer Häufigkeit (über 30% aller Präpositionalphrasen mit *“von”*!) eine explizite Kodierung legitimieren. Satz 37a enthält eines der wenigen stereotypen Sprachmuster, das in der IT festzumachen ist. Folglich enthält die Restriktionsliste der Präposition *“von”* einen spezifischen Constraint COMPANY-PRODUCT, der alle Aktivitäten von Firmen mit Produkten subsumiert. Einen analogen Fall in der Medizin stellt der Satz 37b dar. Beinahe 50% der Okkurrenzen mit der Präposition *“von”* beziehen sich auf die Verknüpfung einer anatomischen Struktur (*“Magenschleimhaut”*) mit einer Typangabe (*“Antrumtyp”*). Auch hier rechtfertigt die hohe Frequenz des Phänomens die Aufnahme einer spezifischeren Relation, HAS-TYPE, in die Restriktionsliste.

37. a. Das Notebook kommt mit dem Cardbus-Controller von Texas Instrument.
(30/97)
- b. Erfäßt wurde eine Magenschleimhaut vom Antrumtyp. (48/103)

Wegen des starken technischen Bezugs und des Leserinteresses an Daten und Preisen finden sich in der IT häufig (mehr als 25% bezogen auf die Präposition *“von”*) Verbindungen von DEGREE-Konzepten mit sie weiter qualifizierenden Maßangaben. Satz 38 beispielsweise kombiniert einen *“Preis”* mit einer Maßeinheit (*“DM”*). Die dazugehörige Relation in der Restriktionsliste lautet HAS-UNIT. In den in der empirischen Studie betrachteten Texten aus der Medizin fehlen Fälle dieser Art vollständig.

38. Der Drucker ist zu einem Preis von 1.200 DM erhältlich. (25/97)

Auch das in Satz 39 anzutreffende Beispiel kommt ausschließlich in der IT vor. Die *“von”*-Präpositionalphrase verbindet eine Konzepteigenschaft (*“Akkulaufzeit”*) mit einem Konzept (*“Toshiba-Notebooks”*). Da der Kopf der Präpositionalphrase die Eigenschaft anzeigt, enthält die Restriktionsliste die inverse Relation ATTRIBUTE-OF. Daraus wird auch

ersichtlich, daß dieser Typ von Präpositionalphrase als Inversion einer “mit”-Präpositionalphrase angesehen werden kann: “*Akkulaufzeit von Notebooks ist befriedigend*” $\Leftrightarrow$ “*Notebooks mit befriedigender Akkulaufzeit*”.

39. Die Akkulaufzeit von Toshiba-Notebooks ist befriedigend. (12/97)

Die nächste Klasse der “von”-Präpositionalphrasen ist insofern interessant, als sie als sinnverwandte Paraphrase eines Genitivattributs mit Verbnominalisierung angesehen werden kann. Somit liefert sie weitere empirische Evidenzen für die bereits bei Genitivattributen diskutierte Subklassenbildung im Lexikon, weil dieselben Interpretationsrestriktionen greifen. Während in der IT nur etwas mehr als 10% der Okkurrenzen dieser Kategorie angehören, erhöht sich deren Anteil in der Medizin auf immerhin knapp 30%.

40. a. Die Unterstützung von Fax-Class ist gewährleistet. (11/97)

b. Der Befund entspricht einer Antrumgastritis mit Nachweis von *Helicobacter pylori*. (30/103)

Im Satz 40a enthält die Präpositionalphrase ein Präpositionalobjekt “*Fax-Class*”, das entweder als *subject* oder als *direct-object* der Verbkonzepts SUPPORT (für “*Unterstützung*”) zu interpretieren ist. Der entsprechende Constraint ist an der Wortklasse für Verbnominalisierungen kodiert. Ob der von der Präposition abhängende Modifikator im Sinn eines Subjekts oder eines direkten Objekts zu interpretieren ist, muß semantisch entschieden werden; oftmals liefert die semantische Interpretation ein ambiges Ergebnis. Der aus der Medizin stammende Satz 40b wird analog behandelt, da es sich bei der Präpositionalphrase auch in diesem Beispiel um ein Präpositionalobjekt handelt. Im Gegensatz zum IT-Beispiel ergibt die Interpretation keine Ambiguität, da das Weltwissen als AGENT von Nachweisen (PROVIDE-EVIDENCE) eine PERSON als Typ-Constraint festlegt. Ohne Zweifel subsumiert das Konzept PERSON *HELICOBACTER-PYLORI* nicht. Deshalb bleibt nur die PATIENT-Lesart als gültige Interpretation.

Als zahlenmäßig eher von geringerem Gewicht zeigt sich die Interpretation der Präpositionalphrasen mit “von” als SOURCE. Gerade mal sechs Okkurrenzen in der IT und vier in der Medizin lassen sich festmachen. Das IT-Beispiel in Satz 41a benennt “*Intel*” als die Quelle eines RECEIVE-Ereignisses. Im Satz 41b aus der Medizin führt die Präpositionalphrase den lokalisatorischen Ursprung eines Tumors (“*ein vom Magen ausgehender Tumor*”) ein.

41. a. Der Anwender bekommt von Intel Motherboards mit dem Triton-Chipsatz.
(6/97)
- b. Es handelt sich um einen vom Magen ausgehenden Tumor. (4/103)

Von ebenfalls geringer Häufigkeit sind Sätze, in denen ein Präpositionalobjekt an ein Verb angebunden wird. Wie bereits für die analogen Beispiele für die Präposition *“mit”* erläutert, läuft die semantische Interpretation über eine am Verb selbst spezifizierte dedizierte Valenz. Dementsprechend wird die Präpositionalphrase *“von 2D-Geschwindigkeit”* als *direct-object* des Verbs *schwärmen* interpretiert. Die semantische Interpretation der Präpositionalphrase *“von der Probe”* in Satz 42b erfolgt demgegenüber als *indirect-object*, da das Verb *“angefertigt”* einen anderen Valenzrahmen aufspannt, jedoch führt auch diese Phrase ein Präpositionalobjekt ein.

42. a. Der Nutzer schwärmt von 2D-Geschwindigkeit. (5/97)
- b. Von der Probe werden weitere Schnittproben angefertigt. (4/103)

Eine besondere Funktion haben *“von”*-Präpositionalphrasen in Passivsätzen, weil sie das syntaktische Subjekt des aktivischen Satzes enthalten können. Liegt eine derartige Konstruktion vor, dann sind die Interpretationsrestriktionen auf *subject* festgelegt. An der Wortklasse des passivischen Partizips existiert die ausgezeichnete Valenz *ppsubject*, welche diesen Constraint beinhaltet. Beispiele für Passiv-Sätze finden sich in 43a und 43b. Die Medizin verwendet häufiger das Passiv, worin sich möglicherweise eine stilistische Präferenz ausdrückt.

43. a. Der Grafikspeicher wird von den Herstellern nicht genutzt. (4/97)
- b. Die Foveolen sind von Epithelien gesäumt. (17/103)

Das letzte mit der Präposition *“von”* verbundene und auch recht seltene Phänomen betrifft die Angabe einer unteren Grenze in einem Intervall, das sich in der IT – wie der Satz 44a veranschaulicht – auch auf Versionsnummern einer Software oder Ähnliches beziehen kann. Meistens tritt die Angabe einer unteren Grenze zusammen mit einer oberen Grenze auf (*“auf OS 2.0”*). Im Beispielsatz aus der Medizin gibt die Präpositionalphrase *“von 45 cm”* die Lage des Tumors im Gastrointestinaltrakt an, was angesichts der extremen sprachlichen Verkürzung des objektiven Sachverhalts nicht richtig interpretiert werden kann, da

Präposition	richtig (eindeutig)	richtig (mehrdeutig)	falsch	NIL	ohne	gesamt
bei (IT)	8 (32%)	0,5	7,5 (30%)	2	7	25
bei (Medizin)	2,5 (23%)	2	4,5 (41%)	2	–	11
mit (IT)	71 (53%)	12,5	13,5	18	20	135
mit (Medizin)	106 (62%)	14	21,5	27,5	1	170
von (IT)	80 (82%)	4	0	3	10	97
von (Medizin)	79 (77%)	4	1	17	2	103

Tabelle 6.7: Verteilung der Interpretationsergebnisse auf Präpositionen

durch die Verbindung mit dem syntaktischen Kopf “*Tumor*” als physikalischem Objekt eine Größenangabe vermutet wird.

44. a. Die Aufrüstung von OS 1.3 auf OS 2.0. (4/97)
- b. Tumor von 45 bis 55 cm. (1/103)

Eine abschließende Betrachtung der Präpositionalphrasen mit “*von*” ergibt ebenfalls eine starke Fokussierung der Medizin auf im Großen und Ganzen drei Interpretationsklassen: die sehr spezifische Typzuordnung zu anatomischen Objekten, die Interpretation von Verbnominalisierungen sowie Passiv-Sätze (95 von 103 Okkurrenzen). Die IT weist untypischerweise ein sehr spezielles Interpretationsschema zwischen Herstellern und Produkten auf, das sich allerdings durch die Textauswahl, nur Produktberichte zu betrachten, begründen läßt.

Zum Abschluß dieses Unterkapitels über die semantische Interpretation von Präpositionalphrasen resümiert die Tabelle 6.7 die Interpretationsergebnisse im IT- und Medizin-Trainingsset, aufgelistet für drei Präpositionen, die beiden ausführlich besprochenen Präpositionen “*mit*” und “*von*” sowie die oftmals Diskursrelationen einführende Präposition “*bei*”. Aus den Daten der Tabelle geht hervor, daß die Interpretationsergebnisse keineswegs prozentual gleichverteilt ist. Die Präposition “*bei*” liefert – bei einer allerdings nur sehr kleinen Zahl an Okkurrenzen – die schlechtesten Ergebnisse. Nur etwa ein Drittel der Vorkommen konnte richtig interpretiert werden. In den übrigen Fällen berechnete die semantische Interpretation entweder ein falsches oder ein *NIL-Ergebnis*, oder es fand keine Interpretation statt.

	GenAtt	ModAux	von	in	für	Summe
IT-Test mit Interpretation	12	45	29	32	22	140
... in IT-Training	2 (13 %)	9 (20%)	11 (38%)	3 (9%)	6 (27%)	31 (22%)
Med-Test mit Interpretation	86	52	41	67	25	271
... in Med-Training	49 (60%)	24 (46%)	31 (76%)	55 (82%)	23 (92%)	182 (67%)

Tabelle 6.8: Stereotypizität des Sprachgebrauchs

Im Vergleich dazu schneidet die Präposition *“mit”* deutlich besser ab. In beiden Domänen überschreitet die Prozentzahl richtiger Interpretationen die 60%-Marke. Die besten Resultate erzielt indessen die Präposition *“von”*, deren Prozentzahl erfolgreicher Interpretationen sich in IT und Medizin immerhin um die 80% bewegt. Die Gründe für die doch auffälligen Unterschiede bei der Verstehensleistung sind im Vortext bereits angeführt und müssen hier nicht wiederholt werden.

Als Fazit bleibt, daß der vorgestellte Ansatz einer schlanken semantischen Interpretation in der Lage ist, gute Resultate zu liefern, wenn die konzeptuelle Zielstruktur klar bestimmbar ist. Die Fälle, in denen unabhängig von der eigentlichen semantischen Interpretation bereits das Problem auf der Ebene der Wissensmodellierung auftritt, sind für die meisten Fehler bzw. nicht ausführbaren Interpretationen verantwortlich.

Weiterhin verursacht offensichtlich eine Reihe von Präpositionalphrasen Probleme, weil sie eine andere Interpretationsschicht adressieren. Sie fungieren im Text entweder ausschließlich oder auch als Indikatoren für Kohärenzrelationen. Diejenigen Präpositionalphrasen, die exklusiv auf außerhalb der eigentlichen Ontologie liegende Kohärenzrelationen abzubilden sind, fallen aus dem Ansatz zur semantischen Interpretation gänzlich heraus, da sie eine andere Ebene der textuellen Struktur berühren. Folglich sind auch andere Interpretationsmodule für deren adäquate Berechnung zuständig (Schauer, 2000).

6.6 Abschließende Bewertung zur Evaluation

Die in der empirischen Untersuchung verwendeten medizinischen Texte sind – im Gegensatz zu den Zeitschriftenartikeln aus der IT – Gebrauchstexte, die massenhaft im klinischen Alltag angefertigt werden. Dadurch sind stilistische Variationen, die einen Text bei der Lektüre interessanter machen, überflüssig. Vielmehr wird ein komprimierter, informativer und ste-

reotyper Sprachstil angewandt. Um zu überprüfen, inwieweit sich die Stereotypizität auch in den Testdaten widerspiegelt, gibt die Tabelle 6.8 für die Genitivattribute, die Modal- und Auxiliärverben sowie eine kleine Anzahl ausgewählter Präpositionalphrasen wieder, wieviele der semantischen Interpretationen im Testset bereits im Trainingsset enthalten sind. Eine semantische Interpretation ist dann auch im Trainingsset enthalten, wenn sie in allen Parametern identisch ist.

Bei den Genitivattributen, für die in der IT die Anzahl allerdings sehr gering ausfällt, finden sich in der IT zwei der insgesamt 12 Interpretationen des Testsets, die bereits im Trainingsset vorkommen (13%). In der Medizin fällt die Übereinstimmung mit 49 von 86 Interpretationen (60%) sehr viel deutlicher aus.

Die Anzahl der übereinstimmenden semantischen Interpretationen erhöht sich unwesentlich bei Modal- und Auxiliärverbkonstruktionen in der IT gegenüber den Genitivattributen. Hier sind 9 von 45 Interpretationen bereits im Trainingsset anzutreffen (20%), was ein Indikator dafür sein könnte, daß Produktbeschreibungen auch über die Verwendung unterschiedlicher Verben sprachlich variieren. Die medizinische Domäne weist indessen wieder eine weitaus stärkere Übereinstimmung auf, 24 von 52 (46%) der Interpretationen wiederholen sich. Dennoch zeigen Modal- und Auxiliärkonstruktionen in der Medizin den vergleichsweise geringsten Grad an Stereotypizität.

Die Präpositionen *“von”*, *“in”* und *“für”* stehen repräsentativ für Stereotypizitätsverhältnisse bei Präpositionalphrasen (vgl. hierzu auch die Aussagen in Kapitel 6.5.2.1). Nur im Fall der Präposition *“von”* gibt es in der IT ein stereotypes, d.h. ein hochfrequentes Präpositionalgefüge des Grundmusters *“Produkt von Hersteller”*. In der Medizin präsentieren sich Präpositionalphrasen mit *für* als hochgradig stereotyp, weil nahezu jeder Befundbericht den Passus *“kein Anhalt für Malignität”* enthält. Präpositionalphrasen mit *in* sind in der IT überhaupt nicht, in der Medizin indessen stark stereotyp.

Vergleicht man den Mittelwert über alle Sprachphänomene, dann kommt die Medizin auf 67% an Wiederholungen in den Anfragen, die IT auf gerade mal 22%, wobei dieser Wert dadurch, daß die einzige stereotype Konstruktion in der IT durch die Präposition *“von”* miteinbezogen ist, eine Abschätzung nach oben hin abgibt.

Als Fazit der kleinen Untersuchung zu stereotypem Sprachgebrauch kann gesagt werden, daß es in der Medizin möglich ist, in eingeschränkten Domänen durch systematische Analyse der syntaktisch realisierten Verbindung von Konzepten zu einer nahezu vollständigen Abdeckung zu kommen. Dagegen bieten Zeitschriftentexte aus der Informationstech-

nologie diese Möglichkeit nicht, weil den wenigen sprachlichen Stereotypen zum Trotz abwechslungsreiches Schreiben zum beruflichen Handwerk gehört. Aus diesem Grund treten deutlich häufiger noch nicht in die Modellierung einbezogene Konstruktionen auf.

Vergegenwärtigt man sich in einer Gesamtbewertung die Resultate aller Evaluationen für Trainings- und Testset in beiden Domänen, dann bleibt festzuhalten, daß die Ergebnisse, die durch die Evaluationsmaße Recall und Precision wiedergegeben sind, sehr ermutigend ausfallen. Behält man indessen die Gesamtverstehensleistung bei der Textanalyse im Auge, dann ist auch zu beachten, daß der kompositionale Aufbau von Bedeutungsstrukturen diese Teilergebnisse zusammenfügt. Gemessen an komplexen, d.h. aus mehreren Phrasen oder gar Sätzen gewonnenen Interpretationsergebnissen, werden sich die Resultate mit Sicherheit verschlechtern. Dennoch ist es methodisch notwendig, zuerst die elementaren Verstehensschritte zu evaluieren, bevor komplexere untersucht werden. Nur auf diese Weise können Fehlerquellen erkannt und Leistungen einzelner Module zur Bedeutungsgewinnung genau ermittelt werden. Genau und nur hier liegt der Beitrag dieser empirischen Studie zur semantischen Interpretation.

Kapitel 7

Ausblick auf die weitere Entwicklung von SYNDIKATE

Der nachfolgende kleine Ausblick erhebt keinesfalls den Anspruch, vollständig die möglichen Arbeiten zu qualitativen und quantitativen Erweiterungen von SYNDIKATE aufzuzählen. Vielmehr resümieren die angeführten Punkte die Erfahrungen, die beim Umgang mit den Texten und mit dem System gemacht wurden. Auf der Grundlage dieser Erfahrungen lassen sich Vorschläge machen, die eine gewisse Priorität für die weitere Systementwicklung genießen. Ausschließlich diese sind im folgenden kurz angeführt.

7.1 Desiderata des Textverstehenssystems SYNDIKATE

Aufgrund der Wissensbasierung nimmt die fortlaufende qualitative und quantitative Erweiterung des Domänenwissens einen hohen Stellenwert ein. Eine substantielle Vergrößerung der Domänenwissensbasen zu IT und Medizin, aber auch der domänenunabhängigen Upper-level-Ontologie kann nur dadurch geleistet werden, daß andere Wissensrepositorien erschlossen werden. Die qualitative Verbesserung der Ontologien ergibt sich durch die empirische Auseinandersetzung mit den Textkorpora, da sich aus den Verwendungskontexten von Konzepten typische Relationierungen ermitteln lassen, anhand derer die Konzept- und Rollendefinitionen – vor allem dank des Path Finders – einer fortlaufenden Prüfung unterworfen sind.

Neben den Erweiterungen in den Wissensbasen gibt es auch noch Potentiale für Verbesserungen am Parser. An vorderster Stelle steht hier die Integration der Koordination mit den dazugehörigen grammatikalischen Spezifikationen.

7.2 Verbesserung und Erweiterung der semantischen Interpretation

Der eingeführte Ansatz zur semantischen Interpretation beschreibt ein schlankes, wohl-definiertes und verallgemeinerbares Verfahren, mit dem die über syntaktische Strukturen verbundenen Diskursentitäten eines Texts konzeptuell relationiert werden. Obwohl bereits eine überzeugende Abdeckung bezüglich der in Texten vorkommenden Konfigurationen erreicht ist, gibt es eine Reihe von Phänomenen, die noch nicht interpretierbar sind.

Als aus linguistischer Perspektive stärkstes Manko erweist es sich in diesem Zusammenhang, daß es beim momentanen Stand der Implementation weder eine Möglichkeit zur Interpretation noch zur Repräsentation von Quantifikation gibt. Im Horizont der Anwendungsorientierung wäre jedoch zunächst empirisch zu bestimmen, wie hoch der Stellenwert der Quantifikation bei auf Textwissensbasen aufsetzenden Anwendungen denn tatsächlich ist. Unabhängig vom Ergebnis dieser Untersuchung sollten prinzipielle methodische Überlegungen im durch SYNDIKATE gegebenen Kontext angestellt werden, um die unterschiedlichen Möglichkeiten zu klären, wie sich die Behandlung der Quantifikation in die Systemarchitektur mit allen ihren Beschränkungen einfügt.

In der medizinischen Domäne ist die semantische Verarbeitung der expliziten (*“kein Anhalt für”*) wie auch der impliziten (*“Hinweise auf eine bakterielle Infektion fehlen”*) Negation von größerer Bedeutung. Der Aussagewert von Befunden, Beobachtungen und Symptomen hängt stark vom bejahenden oder negierenden Kontext ab.

7.3 Erweiterung und Verfeinerung der empirischen Arbeiten

Obwohl sich die empirischen Arbeiten über die semantische Interpretation bewußt auf ein einziges Verarbeitungsmodul des Textverstehenssystems beschränken, bleibt das Problem bestehen, daß die Interpretationsleistung durch mehrere Faktoren beeinflusst wird.

Zunächst bestimmt die statische Interpretation, die konzeptuelle Beschränkungen an Abhängigkeitsrelationen und lexikalische Knoten knüpft, die Menge der Ergebnisse. Es stellt sich die Frage, ob die gefundenen Beschränkungen eng genug gefaßt sind, um das Anwachsen von Ambiguitäten bei der semantischen Interpretation zu begrenzen, und ob sie weit genug gefaßt sind, um die richtigen semantischen Interpretationen tatsächlich zu berechnen. Auf der Seite der Wissensbasis interferieren die konzeptuellen Beschränkungen mit einer adäquaten Abstraktion in der Relationshierarchie. Die Relationshierarchie hat den Zweck, sehr spezielle Typbeschränkungen von Rollenfüllern in Konzeptdefinitionen zu ermöglichen. Dabei muß jede interpretationsrelevante Relation von den richtigen abstrakten Relationen an der Spitze der Relationshierarchie subsumiert werden, damit sie von der semantischen Interpretation angesprochen werden kann.

Die Menge der definierten Relationen hängt überdies mit der Granularität in der Wissensbasis zusammen, da die Konzeptbeschreibungen der Wissensbasis durch eine große Anzahl sehr spezieller Relationen entsprechend viele Relationen aufweisen. Die Menge an Konzepten bestimmt in Verbindung mit den sie definierenden Relationen die Länge der berechneten Pfadmuster, die im Laufe der dynamischen Interpretation berechnet werden.

Um zu einer adäquaten Evaluation der Teilkomponente zur semantischen Interpretation zu kommen, wäre eine parametrisierbare Evaluationsumgebung notwendig, mittels derer die unterschiedlichen Effekte mit Hilfe veränderbarer Parameter eruierbar wären (vgl. Friedman et al. (1998)). Zugleich ergäbe sich der Nutzen, daß praktische Hinweise für Entwurfs- bzw. Modellierungsprinzipien gewonnen würden. Zu variieren ist beispielsweise die maximale Länge der Pfade zwischen Konzepten. Darüber hinaus könnte auch getestet werden, wie sich die Anzahl von Mehrdeutigkeiten im Verhältnis zur Größe der unterliegenden Wissensbasis ändert.

Die modellierten Wissensbasen haben den Zweck, die semantische Interpretation von minimalen Subgraphen (und auch anderen Strukturen) in *Texten* zu unterstützen, d.h. auch die Variationen von Texten zwischen Domänen und innerhalb einer Domäne sind zu untersuchen. Eine äußerst spannende Fragestellung betrifft die Verwendung und die Häufigkeit sprachlicher Mittel in Abhängigkeit von der Textsorte und des Textgegenstands. Zum Beispiel scheinen Genitivattribute in den verwendeten, sehr kurzen Produktbeschreibungen deutlich seltener vorzukommen als in längeren Texten. In den hier verwendeten medizinischen Texten (histopathologische Befundberichte) treten bestimmte semantische Muster besonders häufig auf (z.B. *“in”* nur in seiner lokalen Lesart). Verhält es sich in anderen

Texten der gleichen Domäne, beispielsweise Lehrbuchtexten, genauso?

Die vielen Möglichkeiten, das Testszenario zu variieren, weisen darauf hin, daß sehr viel mehr empirische Arbeiten notwendig sind, um die einzelnen das Textverstehen beeinflussenden Faktoren zu isolieren und deren Bedeutung im Verstehensprozeß bewerten zu können.

Kapitel 8

Zusammenfassung

Diese Arbeit behandelte ein schlankes Verfahren zur semantischen Interpretation im Rahmen des Textverstehenssystems SYNDIKATE. Anstatt im Interpretationsprozeß zu berechnende konzeptuelle Bezüge in einer großen Regelmenge zu kodieren, kommt eine kleine Anzahl (ein knappes Dutzend) von abstrakten semantischen Interpretationsregeln zur Anwendung. Aufbauend auf unterschiedlichen Abstraktionsebenen im Weltwissen und im Lexikon werden die Interpretationsregeln mit Parametern aus dem Interpretationskontext instantiiert. Dabei setzt die semantische Interpretation auf einer Menge von wohldefinierten Konstellationen in Dependenzgraphen auf: den minimalen interpretierbaren Subgraphen, die zwei Inhaltswörter über einen Pfad aus Dependenzrelationen verbinden. Die semantische Interpretation besteht im Kern aus zwei Teilen. Zuerst werden statische Beschränkungen für die Interpretation bestimmt, die sich aus den Dependenzrelationen und den intervenierenden Knoten im minimalen Subgraphen ableiten lassen. Anschließend erfolgt eine dynamische Suche nach konnektierten Pfaden im Konzeptgraphen zwischen Konzepten, die zu den Inhaltswörtern gehören. Diese Suche berücksichtigt die statischen Beschränkungen, damit die Anzahl der gefundenen Pfade begrenzt wird und somit nicht zu viele interpretatorische Ambiguitäten entstehen.

Die angewandte Methode ermöglicht – neben der Ökonomie bei der Regelmenge – eine strikte Trennung der Wissensebenen. Allein durch die statische Interpretation und die Assoziation von Wortaktoren mit Instanzen in der Textwissenbasis sind Syntax und Konzeptsystem aufeinander bezogen. Vermittelt wird zwischen diesen beiden Wissensebenen durch die semantische Ebene, die auf der Grundlage der syntaktischen Strukturen die Instanzen in der Textwissensbasis manipuliert.

Auch die Verwaltung von Ambiguitäten, die im Laufe der Textanalyse entstehen, folgt ganz dem Prinzip, Wissensquellen zu modularisieren. Konzeptuelle und interpretatorische Ambiguitäten werden auf der Seite der Wissensbasis verwaltet, strukturelle Ambiguitäten auf der Seite des Parsers. Dabei bleiben dem Parser die Interpretationsalternativen auf der semantisch-konzeptuellen Seite verborgen. Der Ansatz zur Behandlung von Ambiguitäten, der auf einem Kontextmechanismus basiert, repräsentiert unterschiedliche Lesarten eines Texts in Form disjunkter Mengen von assertionalen Axiomen.

Eingebettet ist der Ansatz zur semantischen Interpretation im Textverstehenssystem SYNDIKATE, dessen Konzeption ausführlich zur Darstellung kam und das in den Zusammenhang von Informationsextraktionssystemen und anderen Textverstehenssystemen gerückt wurde. Insbesondere die Bedeutung resolvierter anaphorischer Phänomene für die Erzeugung valider, kohärenter und vollständiger Textwissensbasen ist verdeutlicht worden.

Da in SYNDIKATE die semantische Interpretation wie auch die Resolution textueller Phänomene wissensbasiert ist, wurde die Wissensrepräsentationsebene von zwei Seiten angegangen. Zum einen sind die Zusammenhänge zwischen Wissensrepräsentationssprachen, Wissensrepräsentationssystemen und Textverstehen aufgezeigt worden. Dabei stellte sich heraus, daß die Beschränkungen aus der Wissensrepräsentation die Tiefe des Verstehens beeinflussen. Insbesondere in Systemen mit größeren Wissensbasen sind grundlegende Entscheidungen zu treffen, welche Anforderungen an das Textverstehen gestellt werden. Zum anderen kamen verschiedene Wissensrepositorien unter der Fragestellung zur Darstellung, wie sie eine ontologisch schlüssige Modellierung und eine Vergrößerung der Wissensbasen zu unterstützen geeignet sind.

Der letzte Teil dieser Arbeit befaßt sich mit einer Evaluierung der semantischen Interpretation. Die Datengrundlage dieser Evaluierung besteht aus medizinischen und informationstechnologischen Texten, die in einem randomisierten Verfahren aus den zur Verfügung stehenden Textkorpora gewonnen wurden. Anschließend erfolgte eine Teilung der Texte in ein Trainingsset und ein Testset. Aus diesen Texten wurden *alle* Genitivattribute, Modal- und Auxiliarkonstruktionen und alle Präpositionalphrasen extrahiert. Die Ergebnisse der Interpretation erreichten gute Recall- und Precisionwerte, wobei die Interpretation der Präpositionalphrasen erwartungsgemäß etwas schlechter ausfiel als die der Genitivattribute und der Modal- bzw. Auxiliärverben. Die Abweichungen bei Precision und Recall zwischen Trainingsset und Testset waren nicht sehr stark. Vor allem die Medizin mit häufig wiederkehrenden Phrasen erreichte nahezu gleichhohe Werte. Insgesamt zeigen die Ergebnisse,

daß die grundlegende Konzeption zur semantischen Interpretation gute Ergebnisse liefert, insbesondere wenn im Weltwissen keine Spezifikationlücken bestehen.

Ein abschließendes starkes Argument für die Verallgemeinerbarkeit des Ansatzes, mit wenigen Interpretationsregeln auskommen zu können, wenn auf allen Wissensebenen von den entsprechenden Abstraktionen Gebrauch gemacht wird, besteht in der Verwendung zweier sehr heterogener Domänen. In der ursprünglichen Version konnte SYNDIKATE nur IT-Texte analysieren. Die Hinzunahme der Medizin machte in der semantischen Ebene nur eine Erweiterung der Restriktionslisten an Präpositionen erforderlich, ansonsten mußten keine weiteren Adaptionen vorgenommen werden.

Anhang A

Aufgabendefinition zur Textklassifikation

Definition der Aufgabe: die in der Textkollektion enthaltenen Artikel aus Computerzeitschriften sind entlang von zwei Dimensionen zu klassifizieren: Textsorte und Thema; hierbei interagieren die beiden Dimensionen nicht. Für jede dieser Dimensionen sind die Kategorien für die Klassifikationsaufgabe vorgegeben. Sie werden in einer natürlichsprachlichen Definition den Probanden zur Verfügung gestellt.¹

Bei der Klassifikation sollte pro Text pro Dimension nur eine Kategorie vergeben werden. Unter keinen Umständen ist die Vergabe von mehr als einer Kategorie zulässig. Allerdings kann im Bedarfsfall das Feld "unsicher" markiert werden, um anzugeben, daß das Klassifikationsergebnis nicht eindeutig ist. Die Texte fokussieren zum Teil nicht nur auf ein Thema, beispielsweise beinhalten Produktberichte über Computer-Systeme (Thema Hardware) auch Bemerkungen über die mitgelieferte Software und das ausgelieferte Betriebssystem oder umgekehrt beinhalten Produktberichte über Bildbearbeitungssoftware (Thema Software) Beschreibungen der Hardwareanforderungen. Nur wenn wirklich kein eindeutigen Themenfokus angegeben werden kann, ist hier die Vergabe von mehr als einer Themakategorie möglich.

Zur weiteren Beachtung: zentral für die Klassifikation sind die Intention des Autors und der Textfokus (welcher bspw. auch durch die Überschriften gesetzt wird.)

¹Die in den Anhängen enthaltenen Definitionen sind die Originaltexte, wie sie den Klassifizierenden ausgehändigt wurden. Auch wenn aus der Erfahrung mit der Klassifikation heraus nun einige Dinge anders formuliert würden, sind die Texte im Original belassen.

Definition der Textsorten:

1. *Teachings*: hierzu gehören Tips & Tricks, How-Tos sowie Tutorials. In diesen Texten wird beschrieben wie man ein konkretes Problem löst oder wie man sich bei einem konkreten Problem verhält. Dies kann sehr verschiedene Themen berühren (z.B. wie verhält man sich, wenn das Finanzamt den Computer nicht als Werbungskosten akzeptiert; was ist zu tun, wenn das CD-Rom alphabetagamma nicht erkannt wird etc.). How-Tos schildern Lösungswege für die Installation von Software und Hardware. Im Gegensatz zu Tips und Tricks gehen sie allgemeiner und systematischer vor und sind nicht auf ein ganz enges Problem eingegrenzt. Tutorials sind (auf möglicherweise mehrere Folgen angelegte) praktische Übungen zur Erweiterung der Kenntnisse in den Gebieten Software, Programmierung, Hardware etc. (Beispiel: Makroprogrammierung in VBA; typisches Merkmal ist hier die Verwendung von Programmcode).
2. *Glossar/Editorial*: Diese Textsorte ist i.a. frei gestaltbar, referenziert auf konkrete Ereignisse innerhalb des Erscheinungskontexts der Zeitschrift und benutzt oftmals Ironie und fiktionale Elemente.
3. *Interview*: Wohl die am leichtesten zu erkennende Textsorte. Klare Struktur durch Frage-Antwort-Interaktion. Je nach redaktioneller Überarbeitung dringt teilweise die mündliche Ausdrucksweise gegenüber der schriftlichen durch.
4. *Belletristik*: zur Belletristik gehören i.a. fiktionale Texte wie kurze Erzählungen, Science Fiction Stories.
5. *Leserbrief*: ein Leserbrief ist Beitrag eines Lesers in Briefform, der manchmal von der Redaktion der jeweiligen Computerzeitschrift beantwortet wird. Auch die Antwort der Redaktion gehört dann zur Textsorte Leserbrief.
6. *Bericht*: im Gegensatz zum Test bewahrt der Bericht eine primär sachliche Form, ohne überwiegend Werturteile über das behandelte Thema abzugeben. Die Hauptintention (das Schwerpunktkriterium) ist das Berichten. Kurztexte mit Wertungen sind per definitionem Berichte, da sie nicht den Platz für einen eigentlichen Text einnehmen. Zum Beispiel: Compaq liefert seine Computer-Systeme neuerdings ausschließlich mit 17-Zoll Monitoren aus; oder: Neues Buch eines Verlags über Java-Programmierung.

7. *Test*: im Gegensatz zum Bericht beinhaltet ein Test wertende Aussagen über sein Thema, welche auf objektivierbaren Kriterien (z.B. standardisierte Testverfahren) beruhen können, jedoch auch subjektive Ansichten des Autors widerspiegeln können. Die primäre Intention bleibt indessen das Werten, auch wenn hierzu Fakten berichtet werden. Bsp.: Im Benchmarktest WinBapco erreichte die Grafiksoftware xyz, bzw., Die Benutzerführung durch das Menü des Grafikprogramms ist unübersichtlich.

Definition der Themen:

1. *Hardware*: Alle Texte, die sich inhaltlich hauptsächlich mit Hardware beschäftigen. Es können einzelne Produkte (z.B. ein Computer-System) oder Klassen von Produkten (z.B. alle IDE-Festplatten) betrachtet werden. Beschreibt ein Text zum Beispiel ein Computersystem und geht kurz auf das Betriebssystem oder die mitgelieferte Software ein, so bleibt dieser Text dennoch ein Hardwaretext.
2. *Software*: Alle Texte, die sich inhaltlich hauptsächlich mit Software beschäftigen. Es können einzelne Produkte (z.B. ein Betriebssystem) oder Klassen von Produkten (z.B. alle Linuxdistributionen) betrachtet werden. Beschreibt ein Text zum Beispiel eine Grafiksoftware und geht kurz auf die Hardwareanforderungen ein, so bleibt dieser Text dennoch ein Softwaretext.
3. *Dienstleistungen/ Dienstleister/ Produktanbieter*: In den Texten geht es um das Anbieten eines Produktes, welches aus einer Dienstleistung besteht, oder aber auch um Unternehmen, die eine Dienstleistung anbieten; dazu gehören beispielsweise Bereitstellung von Internetzugängen, Anbietung von Crossmedia- oder Online-Publikationen, Internetprovider etc.
4. *Recht und juristische Fragen/ Politik*: Diese Texte berühren juristische und politische Fragen der Informationstechnologie und deren konkreten Anwendungen. Zum Beispiel Gültigkeit der AGB's beim Einkaufen im Internet, Verletzung des Markenschutzes von Produkten etc.
5. *Social Events*: Der Text geht auf die Ereignisse in und um Messen, Kongresse, Tagungen etc. ein.

6. *Wirtschaft/ Wirtschaftsnachrichten*: Texte, die sich mit Nachrichten aus der Wirtschaft beschäftigen. Z.B. Absatz von Prozessoren, Roadmap für Prozessorpreise, Firmenfusionen, gesuchte Klientel etc.
7. *Publikationen*: Die Texte befassen sich mit Publikationen oder Informationsdarbietungen unabhängig vom Medium, das diese Publikationen verwenden (Internet, CD-Rom, Buch, aber auch Computerspiele etc.). Im Vordergrund steht die Inhaltsvermittlung, auch wenn sie durch ein Programm gewährleistet wird.
8. *Technologie/ Verfahren/ Normen/ Normierung/ Standardisierung*: Diese Texte beschreiben eine Technologie oder ein Verfahren. Bsp.: neues Verfahren zur Herstellung von Flachbildschirmen oder zur Markteinführung eines Produkts; z.B.: neue TCO99 oder neue Java-API etc.
9. *Sonstige*: Fragen der Gesundheit, usw., jedoch kein Sammelbecken für Mischkategorien!!!

Anhang B

Aufgabendefinition zur Klassifikation von Präpositionalphrasen

Insgesamt fünf Freiburger Studentinnen, die sich zum Zeitpunkt der empirischen Studie im Grundstudium des Fachs Computerlinguistik befanden, nahmen an der Klassifikationsaufgabe zu Präpositionalphrasen teil. Den Studentinnen war weder der Kontext des Textverstehenssystems SYNDIKATE bekannt noch die Aufgabenstellung der Evaluation zur semantischen Interpretation. Im ersten Schritt erstellten sie nach dem Konsensprinzip eine gemeinsame Datengrundlage für die eigentliche Klassifikationsaufgabe. Dazu mußte jede der Studentinnen zunächst alle im Trainings- und Testset der informationstechnologischen Zeitschriftentexte enthaltenen Präpositionalphrasen identifizieren. Anschließend kamen nur solche Präpositionalphrasen in die Klassifikation, deren Modifikator ein Substantiv war. Adverbien, Adjektive und Pronomina schieden aus, da sie momentan nicht Teil der in die empirischen Arbeiten einbezogenen Wortklassen sind.¹ Die verbleibenden Okkurrenzen wurden darauf untersucht, ob sie in ihrem syntaktischen wie semantischen Kontext ambig sind, wobei satzübergreifende Information nicht berücksichtigt, sondern nur satzintern interpretiert wurde.

Nachdem auf diese Art und Weise eine verlässliche Datengrundlage geschaffen wurde, mußten alle Studentinnen entsprechend den im nachfolgenden aufgelisteten Klassifikationskriterien jede Präpositionalphrase einschließlich der möglichen ambigen Lesarten klassifizieren. Sie durften grundsätzlich nur eine Kategorie pro Okkurrenz vergeben.

¹Im Trainingsset reduzierte sich dadurch die Anzahl der betrachteten Präpositionalphrasen um 13 von 685 auf 672.

Die folgende Aufstellung enthält die Definitionen, mit deren Hilfe die Präpositionalphrasen im Text durch die Klassifizierenden zu klassifizieren waren in unveränderter Form.

1. *Funktionsverbgefüge*: Funktionsverben sind die Verben “stehen, stellen, liegen, ziehen, bringen, kommen, versetzen, geraten etc.”, im Kontext einer Reihe von Präpositionalgefügen. Eigenheiten der Funktionsverben sind:
 - (a) Es gibt (fast immer) ein einfaches Verb, welches die gleiche Bedeutung aufweist. (Bsp.: etwas zum Vortrag bringen $\Leftrightarrow$ vortragen; in Betracht ziehen $\Leftrightarrow$ erwägen).
 - (b) Das abhängige Nomen kann nicht weiter modifiziert werden (Bsp.: etwas intensiv in Betracht ziehen (*etwas in intensiven Betracht ziehen).
2. *Konjunktionsähnliche Präpositionalgefüge*: (teilweise koordinierend) Erfüllen im Satz konjunktionsähnliche Funktion, indem sie gleich- oder nebengeordnete Sätze einleiten. (Bsp.: auf der einen Seite, zum einen, zum anderen etc.)
3. *Präpositionsähnliche Nominalgefüge*: Feste Verbindungen, die teilweise schon präpositionalen Charakter aufweisen, d.h., es gibt i.a. zwei Präpositionen im Gefüge, die beide über einen substantivischen Modifikator verfügen. Ähnlich wie bei Funktionsverben wird das Substantiv im Präpositionalgefüge im allgemeinen nicht modifiziert (Bsp.: in Bezug auf, *in schönem Bezug auf*). Einige dieser Gefüge haben den Prozeß der Präpositionswordung bereits weitgehend abgeschlossen (Bsp.: anhand/ an Hand von, anhand des, mit Hilfe von, mit Hilfe/mithilfe des).
4. *Adverbielle Präpositionalphrasen (an Verben und an Adjektiven) bzw. attributive Präpositionalphrasen (an Substantiven)*:

Vorbemerkung: Die meisten Präpositionen führen grundsätzlich nie ein Präpositionalobjekt ein. Aus der großen Anzahl seien ein paar Beispiele herausgegriffen: Postpositionen (zufolge, entsprechend ...), Zirkumpositionen (von ... an, um ... willen, ...), denominalen Präpositionen (kraft, trotz, infolge, aufgrund ...), deverbale Präpositionen (betreffs, entsprechend, ...), Genitiv regierende Präpositionen (längs, jenseits, anlässlich, laut, bezüglich, ...) sowie ab, seit, neben, wider, bis.

Adverbielle Präpositionalphrasen an Verben: Auf der Ebene der Verben geben die adverbiellen Präpositionalphrasen zusätzliche, nicht-obligatorische Informationen zu

Handlungen, Zuständen und Ereignissen, die durch Verben eingeführt sind. Diese zusätzliche Informationen betreffen:

- (a) temporale Ergänzungen
- (b) lokale oder direktionale Ergänzungen
- (c) modale Ergänzungen (Ergänzungen der Art und Weise)
- (d) kausale Ergänzungen (Bsp.: Wegen Guido bin ich gut drauf)
- (e) instrumentale Ergänzungen (Bsp.: Spielen mit einem Fußball)
- (f) komitative Ergänzungen (Bsp.: ich gehe mit Guido in die Stadt)
- (g) konditionale Ergänzungen (Bsp.: bei Regen fühle ich mich wohl)
- (h) finale Ergänzungen (Bsp.: Zur Erholung spiele ich Quake)
- (i) Konstruktionen mit “sein” der Form: in Sicht sein, in Form sein, in Trance sein, an der Reihe sein etc.

Adverbielle Präpositionalphrasen an Adjektiven: Auf der Ebene der Substantive sind alle Präpositionalphrasen adverbial, die nicht Präpositionalobjekte sind. Im übrigen gelten dieselben Kriterien wie bei adverbialen Präpositionalphrasen an Verben. Bsp.: Die nur nach Mitternacht arbeitende Studentin, Guido ist wegen seiner Leistungen stolz.

Attributive Präpositionalphrasen an Substantiven: Auf der Ebene der Substantive sind alle Präpositionalphrasen attributiv, die nicht Präpositionalobjekte sind. Bsp.: Rechner mit Festplatte, das romantische Essen nach Mitternacht, die Freude wegen Guido (aber: “die Freude über Guido” wäre Präpositionalobjekt).

5. *Präpositionalobjekte an Verben, Adjektiven und Substantiven:*

Präpositionalobjekt an Verben: Rein deskriptiv kann nur eine Untermenge von Präpositionen Präpositionalobjekte an Verben (und Kopula plus Adjektiv) einführen. Dazu gehören die folgenden Präpositionen: an, auf, aus, durch, für, gegen, in, mit, nach, über, unter, um, von, vor, zu, hinter, zwischen. Bei Substantiven dürfte es sich ähnlich verhalten. Ein Präpositionalobjekt an Verben liegt vor, wenn:

- (a) es keine adverbielle Präpositionalphrase ist und

- (b) wenn die Präpositionalphrase in der Objektlesart nur einmal zu füllen ist. (Bsp.: Ich arbeite an meiner Dissertation. *Ich arbeite an meiner Dissertation an meiner Magisterarbeit.* Ich arbeite nur an Montagen an meiner Dissertation.)

Weiterhin ist zu beachten, daß die Präpositionalobjekte nicht unbedingt obligatorisch sind. (Bsp.: Ich kümmere mich um meine Mutter (obligatorisch); aber: ich forsche an der Uni (adverbielle Präposition) versus Ich forsche nach Klassifikationskriterien (Präpositionalobjekt)).

Präpositionalobjekte an Adjektiven: Präpositionalphrasen treten an Adjektiven nur in Verbindung mit "sein" auf. (Ausnahme sind die pränominalen Adjektive; Bsp.: Der auf Guido eifersüchtige Martin). Ein Präpositionalobjekt liegt dann vor, wenn

- (a) keine adverbielle Präpositionalphrase vorliegt und
- (b) wenn die Präpositionalphrase in der Objektlesart nur einmal zu füllen ist. (Bsp.: Ich bin stolz auf Guido. *Ich bin stolz auf Guido auf mich*. Auf Anraten meines Bruders bin ich stolz auf Guido.) Weiterhin zu beachten, daß die Präpositionalobjekte nicht unbedingt obligatorisch sind. (Bsp.: Ich bin eifersüchtig versus Ich bin eifersüchtig auf Guido.)

Deverbalisierte Adjektive sind zu klassifizieren wie die korrespondierende Verbkonstruktion. Bsp.: Der nach Mitternacht forschende Professor ... (adverbiale Präposition) Der nach der Mitternachtssonne forschende Professor. (Präpositionalobjekt)

Präpositionalobjekte an Substantiven: Ein Präpositionalobjekt an Substantiven liegt vor, wenn:

- (a) das Substantiv von einem Verb abgeleitet werden kann und
- (b) die vom Substantiv abhängige Präpositionalphrase bzgl. des Verbs Präpositionalobjekt oder Kasusobjekt ist. Bsp.: die Abstammung vom Affen $\Leftrightarrow$ vom Affen abstammen; Installation von Windows $\Leftrightarrow$ Man installiert Windows; Installation vom Systemadministrator $\Leftrightarrow$ der Systemadministrator installiert; Bedarf an Rechnern $\Leftrightarrow$ der Rechner bedürfen; die Reue über die Tat $\Leftrightarrow$ die Tat bereuen.

Idiomatische Wendungen mit Verbalgefüge: In diesen syntaktischen Konstruktionen tritt ein Verb zusammen mit einer feststehenden, möglicherweise modifizierten Nominalphrase auf (Bsp.: ein Gespräch führen, ein langes Gespräch führen, die Flucht

ergreifen, Angst haben, große Angst haben, Maßnahmen ergreifen, drakonische Maßnahmen ergreifen etc.). Die Nominalphrasen können durch eine ein Präpositionalobjekt einführende Präpositionalphrase weiter modifiziert werden. Die Präposition in diesen Konstruktionen ist festgelegt (Bsp.: ein Gespräch führen über etwas/jmd., die Flucht ergreifen vor etwas/jmd., Angst haben vor etwas/jmd., Maßnahmen ergreifen für/gegen etwas/jmd.; bei “Maßnahmen ergreifen” steht entweder “gegen” oder “für”, aber nicht beides).

Die eigentliche Klassifikation der Präpositionalphrasen erfolgt anhand eines festvorgegebenen Entscheidungsbaums. Nach dieser Festlegung ist zuerst zu prüfen, ob ein *Funktionsverbgefüge* vorliegt, wenn dies nicht der Fall ist, dann ob ein *konjunktionsähnliches Präpositionalgefüge* sowie schließlich ob ein *präpositionsähnliches Nominalgefüge* vorliegt. Wenn keine der drei angeführten Klassifikationen gegeben ist, untersucht die letzte Prüfung, ob es sich um eine *adverbielle Präpositionalphrase* (bei Verben und Adjektiven als syntaktischem Kopf) bzw. um eine *attributive Präpositionalphrase* (bei Substantiven als syntaktischem Kopf) handelt oder um ein *Präpositionalobjekt*. Da die beiden Klassen komplementär zueinander sind, läuft die Klassifikation im gleichen Klassifikationsschritt.

Anhang C

Beispieltexte aus dem Trainingsset

Canon: Fotodruck par excellence (Informationstechnologie)

In den USA hat Canon Farbdruker angekündigt, die gleich mit einer ganzen Reihe von Superlativen aufwarten: Der BJC-7000 druckt mit sieben Grundfarben; seine mehr als 1000 Düsen produzieren Tintentropfen in zwei verschiedenen Größen. Um auch auf Normalpapier eine maximale Qualität zu erreichen, sprüht der 1200 mal 600-dpi-Drucker eine transparente Grundierungsschicht aufs Blatt. In den USA soll das Gerät ab Oktober für 450 Dollar erhältlich sein. Preis und Lieferdatum für Deutschland sind noch geheim.

Ebenfalls mit Tintentropfen in zwei verschiedenen Größen arbeitet ein neuer 720 mal 360-dpi-Vierfarbdrucker, der in den USA unter der Bezeichnung 'BJC-4304 Photo' zum Preis von 240 Dollar erhältlich ist. Zum Aufpreis von 99 Dollar verwandelt sich dieser Drucker in einen 24-Bit-Scanner (360 dpi) - eine statt des Druckkopfes montierte Scanneinheit macht's möglich. Ein großer Bruder dieses Druckers für DIN-A3-Großformat wird vermutlich auch über eine Macintosh-Schnittstelle verfügen.

Canon stellt auf der IFA in Berlin aus: Halle 4.1/8

(aus ct September 1997)

Histologisches Gutachten (Medizin)

1. "Ulcus ventrikuli": Zwei Biopsien zeigen zum einen eine Gewebenekrose, zum anderen eine Magenmukosa vom Antrum bzw. Intermediärtyp mit ausgeprägter intestinaler Metaplasie und sehr dichter lymphoplasmazellulärer und granulozytärer entzündlicher Infiltration wobei die Granulozyten auch in das Drüsenepithel einwandern und intraluminal liegen. Des weiteren sind lymphoide Aggregate ausgebildet (1). 2. "Corpus": Das 2 mm große Partikel einer Magenmukosa vom Corpustyp schließt in der Lamina propria mäßig dichte

Lymphozyten und Plasmazellen ein wobei lymphoide Aggregate auftreten. Granulozyten sind nur vereinzelt nachweisbar. Die Befunde entsprechen 1. einem Ulcus mit ausgeprägter chronischer und aktiver Gastritis der Antrummukosa mit intestinaler Metaplasie und lymphoiden Infiltraten, 2. einer mittelgradigen chronischen und gering aktiven Corpusgastritis mit lymphoiden Infiltraten. Über das Ergebnis von Bakterienfärbungen wird getrennt berichtet (2). Kein Anhalt für Malignität.

Literaturverzeichnis

- Agha, Gul, Ian A. Mason, S. F. Smith & C. L. Talcott (1997). A foundation for actor computation. *Journal of Functional Programming*, 7(1):1–72.
- Akman, Varol & Mehmet Surav (1996). Steps toward formalizing context. *AI Magazine*, 17:55–72.
- Ali, Syed S. & Stuart C. Shapiro (1993). Natural language processing using propositional semantic networks with structured variables. *International Journal of Minds and Machines*, 3(4):421–452.
- Allen, James (1993). Natural language, knowledge representation, and logical form. In Madeleine Bates & Ralph M. Weischedel (Eds.), *Challenges in Natural Language Processing*, pp. 146–175. Cambridge: Cambridge University Press.
- Allen, James (1997). *Natural Language Understanding* (2nd ed.). Redwood City, CA: Benjamin/Cummings.
- Allgayer, Jürgen (1990). SB-ONE+: dealing with sets efficiently. In Luiga C. Aiello (Ed.), *ECAI'90 – Proc. of the 9th European Conference on Artificial Intelligence; Stockholm, Sweden, 6–10 Aug 1990*, pp. 13–18. London: Pitman.
- Allgayer, Jürgen. & Carola Reddig (1990). What's in a det? steps towards determiner-dependent inferencing. In Philippe Jorrand & V. Sgurev (Eds.), *AIMSA'90 – Proc. of the 4th International Conference on Artificial Intelligence; Albena, Bulgaria, 19–22 Sep 1990*, pp. 319–328. Amsterdam : North-Holland.
- Appelt, Douglas E., Jerry R. Hobbs, John Bear, David Israel & Mabry Tyson (1993). FASTUS: a finite-state processor for information extraction from real-world text. In

- IJCAI'93 – Proc. of the 13th International Joint Conference on Artificial Intelligence; Chambéry, France, 28 Aug–3 Sep 1993*, pp. 1172–1178. San Mateo, CA: M. Kaufmann.
- Aronson, Alan R., Thomas C. Rindflesch & A. C. Browne (1994). Exploiting a large thesaurus for information retrieval. In *RIAO'94 – Proc. of the 1994 Conference of the Recherche d'Informations Assitee par Ordinateur*, pp. 197–216. Paris: CID.
- Artale, Alessandro & Enrico Franconi (1994). A computational account for description logic of time and action. In *KR'94 – Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning*, pp. 3–14. San Francisco, CA: M. Kaufmann.
- Artale, Alessandro & Enrico Franconi (1998). A temporal description logic for reasoning about actions and plans. *Journal of Artificial Intelligence Research*, 9:463–506.
- Artale, Alessandro, Enrico Franconi, Nicola Guarino & Luca Pazzi (1996). Part-whole relations in object-centered systems: an overview. *Data and Knowledge Engineering*, 20(3):347–383.
- Azzam, Saliha (1994). Classification of noun phrases into concepts or individuals. In *AAAI'94 – Proc. of the 12th National Conference on Artificial Intelligence*, p. 1425. Menlo Park, CA: AAAI Press.
- Baader, Franz (1990). Terminological cycles in KL-ONE-based knowledge representation languages. In *AAAI'90 – Proc. of the 8th National Conference on Artificial Intelligence*, pp. 621–626. Menlo Park, CA: AAAI Press.
- Baader, Franz (1991). Augmenting concept languages by transitive closure of roles: an alternative to terminological cycles. In J. Mylopoulos & R. Reiter (Eds.), *IJCAI'91 – Proc. of the 12th International Joint Conference on Artificial Intelligence; Sydney, Australia, 24–30 Aug 1991*, pp. 446–451. San Mateo, CA: M. Kaufmann.
- Baader, Franz, Hans-Jürgen Bürckert, Bernhard Hollunder, Armin Laux & Werner Nutt (1992). Terminologische Logiken. *Künstliche Intelligenz*, 6(3):23–33.
- Ballmer, Thomas T. & Waltraud Brennenstuhl (1986). *Deutsche Verben. Eine sprachanalytische Untersuchung des deutschen Verbwortschatzes*. Tübingen: G. Narr.

- Barnett, Jim, Kevin Knight, Inderjeet Mani & Elanie Rich (1990). Knowledge and natural language processing. *Communications of the ACM*, 33(8):50–71.
- Barwise, John (1987). Noun phrases, generalized quantifiers and anaphors. In Peter Gärdenfors (Ed.), *Generalized Quantifiers*, pp. 1–29. Dordrecht: Reidel.
- Barwise, John & Robin Cooper (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219.
- Bateman, John A., Robert T. Kasper, Johanna D. Moore & Richard A. Whitney (1990). *A general organization of knowledge for natural language processing: The PENMAN upper model*. Technical Report ISI/RS-87: Univ. of Southern California.
- Bateman, John A., Bernardo Magnini & G. Fabris (1995). The Generalized Upper Model knowledge base: Organization and use. In Nicolaas J. I. Mars (Ed.), *Towards Very Large Knowledge Bases*, pp. 60–72. Amsterdam: IOS Press.
- Baud, Robert H., Anne-Marie Rassinoux, Christian Lovis, Judith C. Wagner, Vincent Griesser, Pierre-Andre Michel & Jean-Raoul Scherrer (1996). Knowledge sources for natural language processing. In James J. Cimino (Ed.), *AMIA'96 – Proc. of the 1996 AMIA Annual Fall Symposium (formerly SCAMC); Washington, D.C., 26–30 Oct, 1998*, pp. 70–74. Philadelphia, PA: Hanley & Belfus.
- Baud, Robert H., Anne-Marie Rassinoux & Jean-Raoul Scherrer (1992). Natural language processing and semantical representation of medical texts. *Methods of Information in Medicine*, 31(2):117–125.
- Baud, Robert H., Anne-Marie Rassinoux, Judith C. Wagner, Christian Lovis, Christian Juge, L. L. Alpay, Pierre-Andre Michel, P. Degoulet & Jean-Raoul Scherrer (1995). Representing clinical narratives using conceptual graphs. *Methods of Information in Medicine*, 34(1-2):176–186.
- Bean, Carol A., Thomas C. Rindflesch & Charles A. Sneiderman (1998). Automatic semantic interpretation of anatomic spatial relationships in clinical text. In Christopher G. Chute (Ed.), *AMIA'98 – Proc. of the 1998 AMIA Annual Fall Symposium; Orlando, FL, 7–11 Nov 1998*, pp. 897–901. Philadelphia, PA: Hanley & Belfus.

- Bernauer, Jochen (1992). Konzeptuelle Graphen als semantisches Modell für den medizinischen Befund. *Biometrie und Informatik in Medizin und Biologie*, 23(4):228–235.
- Bernauer, Jochen (1996). Analysis of part-whole relation and subsumption in the medical domain. *Data and Knowledge Engineering*, 20(3):405–415.
- Biebow, Brigitte & Guy Chaty (1993). A comparison between Conceptual Graphs and KL-ONE. In *ICCS'93 – Proc. of the 1st International Conference on Conceptual Structures*, pp. 75–89. Berlin: Springer.
- Bodenreider, O., A. Burgun, G. Botti, M. Fieschi & P. Le Beux (1998a). Evaluation of the united [sic!] medical language system as medical knowledge source. *Journal of the American Medical Informatics Association*, 5(1):76–87.
- Bodenreider, O., S. Nelson, W. Hole & F. Chang (1998b). Beyond synonymy: Exploiting the UMLS semantics in mapping vocabularies. In Christopher G. Chute (Ed.), *AMIA'98 – Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7–11 Nov 1998*, pp. 815–819. Philadelphia, PA: Hanley & Belfus.
- Bonnema, Remko, Rens Bod & Remko Scha (1997). A DOP model for semantic interpretation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the ACL*, pp. 159–167. Madrid, Spain, July 7–12, 1997.
- Borgida, Alexander (1996). On the relative expressiveness of description logics and predicate logics. *Artificial Intelligence*, 82(5):353–367.
- Bouaud, Jacques, Bruno Bachimont, Jean Charlet & Pierre Zweigenbaum (1995). Methodological principles for structuring an “ontology”. In *Proc. of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing; Montreal, Canada, 19 Aug 1995*. San Mateo, CA: M. Kaufmann.
- Bouaud, Jacques, Bruno Bachimont & Pierre Zweigenbaum (1996). Processing metonymy: a domain-model heuristic graph traversal approach. In *COLING'96 – Proc. of the 16th International Conference on Computational Linguistics; Copenhagen, Denmark, 5–9 Aug, 1996*, pp. 137–142.

- Brachman, Ronald J. (1979). On the epistemological status of semantic networks. In N. Findler (Ed.), *Associative Networks: Representation and Use of Knowledge by Computers*, pp. 3–50. New York: Academic Press.
- Brachman, Ronald J. (1983). What IS-A is and isn't: An analysis of taxonomic links in semantic networks. *IEEE Computer*, 16(10):30–36.
- Briot, Jean-Pierre (1989). ACTALK: a testbed for classifying and designing actor languages in the SMALLTALK-80 environment. In *ECOOOP'89 – Proc. of the European Workshop on Object-Based Concurrent Computing*. Nottingham, U.K., 10–14 Jul 1989, pp. 109–129. Cambridge: Cambridge UP.
- Bröcker, Norbert (1998). Separating surface order and syntactic relations in a dependency grammar. In *COLING'98/-ACL'98 – Proc. of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL; Montreal, Canada, 10–14 Aug 1998*, pp. 174–180.
- Bröcker, Norbert (1999). *Eine Dependenzgrammatik zur Kopplung heterogener Wissensquellen*. Tübingen: M. Niemeyer.
- Bröcker, Norbert, Susanne Schacht, Peter Neuhaus & Udo Hahn (1996). Performanzorientiertes Parsing und Grammatik-Design. Das PARSETALK-System. In Christopher Habel, Siegfried Kanngießer & Gert Rickheit (Eds.), *Perspektiven der kognitiven Linguistik. Modelle und Methoden*, pp. 79–125. Opladen: Westdeutscher Verlag.
- Bröcker, Norbert, Udo Hahn & Susanne Schacht (1994). Concurrent lexicalized dependency parsing: the PARSETALK model. In *COLING'94 - Proc. 15th International Conference on Computational Linguistics*, pp. 379–385.
- Brown, P.J.B., M. O'Neil & C. Price (1998). Semantic definition of disorders in Version 3 of the Read Codes. *Methods of Information in Medicine*, 37(4-5):415–419.
- Buchheit, Martin, Francesco M. Donini, Werner Nutt & Andrea Schaerf (1998). A refined architecture for terminological systems: Terminology = schema + view. *Artificial Intelligence*, 99(2):209–260.
- Buchheit, Martin, Manfred Jesusfeld, Werner Nutt & Martin Staudt (1994). Subsumption between queries to object-oriented databases. *Information Systems*, 19(1):33–54.

- Buvač, Sasa (1995). Resolving lexical ambiguity using a formal theory of context. In *Proc. of the AAAI Fall Symposium on Formalizing Contexts; Cambridge, MA, 10–12 Nov 1995*.
- Buvač, Sasa, Vanja Buvač & Ian A. Mason (1995). Metamathematics of contexts. *Fundamenta Informaticae*, 23(3):263–301.
- Buvač, Sasa & Ian A. Mason (1993). Propositional logic of context. In *AAAI'93 – Proc. 11th National Conference on Artificial Intelligence; Washington, D.C., 11–15 Jul 1993*. Menlo Park, CA; Cambridge, MA: AAAI Press; MIT Press.
- Buvač, Sasa; Buvač, Vanja & Ian A. Mason (1994). The semantics of propositional contexts. In *ISMIS'94 - Proc. of the 8th International Symposium on Methodologies for Intelligent Systems*. Berlin : Springer.
- Campbell, James R., Paul Carpenter, Charles Sneiderman, Simon Cohn, Christopher G. Chute & Judith Warren (1997). Phase II evaluation of clinical coding schemes: Completeness, taxonomy, mapping, definitions and clarity. *Journal of the American Medical Informatics Association*, 4(4):238–251.
- Campbell, K. E., A. K. Das & M. A. Musen (1994). A logical foundation for representation of clinical data. *Journal of the American Medical Informatics Association*, 1(3):218–232.
- Carenini, Giuseppe & J. D. Moore (1994). Using the UMLS semantic network as a basis for constructing a terminological knowledge base: a preliminary report. In C. Safran (Ed.), *SCAMC'93 – Proc. of the 17th Annual Symposium on Computer Applications in Medical Care*, pp. 725–729. New York: McGraw-Hill.
- Carletta, Jean (1997). Assessing agreement on classification tasks: The kappa statistics. *Computational Linguistics*, 22(2):249–254.
- Carpenter, Bob (1992). *The Logic of Typed Feature Structures. With Applications to Unification Grammars, Logic Programs and Constraint Resolution*. Cambridge: Cambridge UP.
- Carpenter, Bob (1997). *Typed-Logical Semantics*. Cambridge, MA: MIT Press.

- CEN/TC251/WG2/PT003 (1995). *Medical informatics - Categorical structures of systems of concepts - Model for representation of semantics. Draft European Prestandard submitted to CEN*. Ref. No. CEN/TC251/PT003/N34 E v4.1d.
- Charniak, Eugene & Robert Goldman (1988). A logic for semantic interpretation. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics; Buffalo, NY, 7-10 Jun 1988*, pp. 87-94.
- Chinchor, Nancy (1998). Overview MUC-7/MET-2. In *MUC-7 – Proc. of the 7th Message Understanding Conference*. San Diego, CA: Saic. [<http://www.muc.saic.com>].
- Cimino, James J. (1995). Coding systems in health care. In J. H. Van Bommel & A. T. McCray (Eds.), *IMIA Yearbook of Medical Informatics'95. The Computer-Based Patient Record*, pp. 71-85. Stuttgart: Schattauer.
- Cimino, James J. (1998a). Auditing the Unified Medical Language System with semantic methods. *Journal of the American Medical Informatics Association*, 5(1):41-51.
- Cimino, James J. (1998b). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5):394-403.
- Cimino, James J. (1998c). Distributed cognition and knowledge-based controlled medical terminologies. *Artificial Intelligence in Medicine*, 12(1):153-168.
- Cimino, James J., Paul D. Clayton, George Hripcsak & Stephen B. Johnson (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association*, 1(1):35-50.
- Coté, R. (1993). *SNOMED International*. College of American Pathologists.
- Creary, Lewis G. & Carl J. Pollard (1985). A computational semantics for natural language. In *ACL'85 – Proc. of the 23rd Annual Meeting of the Association for Computational Linguistics; Chicago, Ill., 8-12 Jul 1985*, pp. 172-179.
- Dahl, Deborah A. (1986). Focusing and reference resolution in PUNDIT. In *AAAI'86 – Proc. of the 5th National Conference on Artificial Intelligence; Philadelphia, PA, 11-15 Aug 1986*, pp. 1083-1088. Los Altos, CA: M. Kaufmann.

- Dahlgren, Kathleen & Joyce McDowell (1986). Using commonsense knowledge to disambiguate prepositional phrase modifiers. In *AAAI'86 – Proc. of the 5th National Conference on Artificial Intelligence; Philadelphia, PA, 11–15 Aug 1986*, pp. 589–593. Los Altos, CA: M. Kaufmann.
- Dalrymple, Mary, John Lamping & Vijay Saraswat (1993). LFG semantics via constraints. In *EACL'93 – Proc. of the 6th Conference of the European Chapter of the ACL; Utrecht, The Netherlands, 21–23 Apr 1993*, pp. 97–105.
- Dichev, Christo (1997). Contexts as communication theories. *AI Communications*, 10(1):185–200.
- Donini, Francesco M., Maurizio Lenzerini, Daniele Nardi & Werner Nutt (1991). Tractable concept languages. In J. Mylopoulos & R. Reiter (Eds.), *IJCAI'91 – Proc. of the 12th International Joint Conference on Artificial Intelligence; Sydney, Australia, 24–30 Aug 1991*, pp. 458–465. San Mateo, CA: M. Kaufmann.
- Donini, Francesco M., Maurizio Lenzerini, Daniele Nardi & Werner Nutt (1997). The complexity of concept languages. *Information and Computation*, 134(1):1–58.
- Donini, Francesco M., Maurizio Lenzerini, Daniele Nardi & Andrea Schaerf (1996). Reasoning in description logics. In Gerhard Brewka (Ed.), *Principles of Knowledge Representation*, pp. 191–235. Stanford, CA: CSLI.
- Dowty, David R. (1986). The effects of aspectual class on temporal structure of discourse: semantics or pragmatics. *Linguistics and Philosophy*, 9:37–61.
- Doyle, John & Ramesh S. Patil (1991). Two theses on knowledge representation: Language restrictions, taxonomic classification, and the utility of representation services. *Artificial Intelligence*, 48(3):261–297.
- Drosdowski, Günther (Ed.) (1984). *Grammatik der Gegenwartssprache* (4 ed.). Mannheim: Duden.
- Dunham, George. (1986). The role of syntax in the sublanguage of medical diagnostic statements. In Ralph Grishman & Richard Kittredge (Eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and processing*, pp. 175–194. London: L. Erlbaum.

- Engel, Ulrich (1988). *Deutsche Grammatik* (2nd ed.). Heidelberg: Groos.
- Fellbaum, Christiane (1998a). A semantic network of English verbs. In Christiane Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 69–104. Cambridge, MA : MIT Press.
- Fellbaum, Christiane (Ed.) (1998b). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Ferrari, Giacomo (1997). Types of contexts and their role in multimodal communication. *Computational Intelligence*, 13(3):414–426.
- Fillmore, Charles J. (1968). The case for case. In Emmon Bach & Robert T. Harms (Eds.), *Universals in Linguistic Theory*, pp. 1–88. New York : Holt, Rinehart & Winston.
- Franconi, Enrico (1993). A treatment of plurals and plural quantifications based on a theory of collections. *International Journal of Minds and Machines*, 3(4):453–474.
- Franconi, Enrico (1994). Description logics for natural language processing. In *Proc. of the 1994 AAAI Fall Symposium on Knowledge Representation for Natural Language Processing in Implemented Systems*, pp. 37–44. New Orleans, Louisiana, November 1994.
- Friedman, Carol (1997). Towards a comprehensive medical language processing system: Methods and issues. In R. Masys (Ed.), *AMIA '97 – Proc. of the 1997 AMIA Annual Fall Symposium (formerly SCAMC); Nashville, TN, 25–29 Oct 1997*, pp. 595–599. Philadelphia, PA: Hanley & Belfus.
- Friedman, Carol, James J. Cimino & Stephen B. Johnson (1994). A schema for representing medical language applied to clinical radiology. *Journal of the American Medical Informatics Association*, 1(3):233–248.
- Friedman, Carol & George Hripcsak (1998). Evaluating natural language processors in the clinical domain. *Methods of Information in Medicine*, 37(4-5):334–344.
- Friedman, Carol, George Hripcsak, William DuMouchel, Stephen B. Johnson & Paul D Clayton (1995a). Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1:83–108.

- Friedman, Carol, George Hripcsak & Irina Shablinsky (1998). An evaluation of natural language processing methodologies. In Christopher G. Chute (Ed.), *AMIA '98 – Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, Florida, 7–11 Nov, 1998*, pp. 855–859. Philadelphia, PA: Hanley & Belfus.
- Friedman, Carol, Stanley M. Huff, William R. Hersh, E. Pattison-Gordon & James J. Cimino (1995b). The Canon group's effort. Working towards a merged model. *Journal of the American Medical Informatics Association*, 2(1):4–18.
- Goldberg, Adele & David Robson (1983). *Smalltalk-80. The Language and its Implementation*. Reading, MA: Addison-Wesley.
- Goldberg, Adele & David Robson (1989). *Smalltalk-80. The Language*. Reading, MA: Addison-Wesley.
- Gomez, Fernando, Carlos Segami & Richard Hull (1997). Determining prepositional attachment, prepositional meaning, verb meaning and thematic roles. *Computational Intelligence*, 13(1):1–31.
- Grishman, Ralph & Lynette Hirschman (1986). PROTEUS and PUNDIT: research in text understanding. *Computational Linguistics*, 12(2):141–145.
- Grishman, Ralph & Beth Sundheim (1996a). Design of the MUC-6 evaluation. In Beth Sundheim (Ed.), *MUC-5 – Proc. of the 5th Message Understanding Conference*, pp. 1–11. San Mateo, CA: M. Kaufmann.
- Grishman, Ralph & Beth Sundheim (1996b). Message understanding conference - 6: A brief history. In *COLING'96 – Proc. of the 16th International Conference on Computational Linguistics; Copenhagen, Denmark, 5–9 Aug 1996*, pp. 466–471.
- Gross, Derek & Katherine Miller (1990). Adjectives in WordNet. *International Journal of Lexicography*, 3(4):265–277.
- Grosz, Barbara J., Avarind K. Joshi & Scott Weinstein (1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Gruber, Thomas R. (1993). Ontolingua: a mechanism to support portable ontologies. *Knowledge Acquisition*, 5:199–220.

- Guha, R. V. & Douglas B. Lenat (1990). CYC: A midterm report. *AI Magazine*, 11(3):32–59.
- Haarslev, Volker, Carsten Lutz & Ralf Möller (1998). Foundations of spatioterminological reasoning with description logics. In Anthony G. Cohn, Lenhart K. Schubert & Stuart C. Shapiro (Eds.), *KR'98 – Proc. of the 6th International Conference on Principles of Knowledge Representation and Reasoning; Trento, Italy, Jun 1998*, pp. 112–123.
- Hahn, Udo, Rainer Hammwöhner, Ulrich Reimer & Ulrich Thiel (1990). Inhaltsorientierte Navigation in automatisch generierten Hypertext-Basen. In P. A. Gloor & N. A. Streitz (Eds.), *Hypertext und Hypermedia - Von theoretischen Konzepten zur praktischen Anwendung*, pp. 205–219. Berlin: Springer.
- Hahn, Udo & Manfred Klenner (1994). Tracking the evolution of concepts in dynamic worlds. In D. Karagiannis (Ed.), *DEXA'94 - Proc. 5th International Conference on Database and Expert Systems Applications*, pp. 410–419. Berlin : Springer. (LNCS, Vol. 856).
- Hahn, Udo, Manfred Klenner & Klemens Schnattinger (1996a). A quality-based terminological reasoning model for text knowledge acquisition. In N. Shadbolt, K. O'Hara & G. Schreiber (Eds.), *EKAW'96 – Proc. of the 9th European Knowledge Acquisition Workshop. Nottingham, U.K., 14–17 May 1996*, pp. 131–146. Berlin: Springer. (LNAI, 1076).
- Hahn, Udo & Katja Markert (1997). In support of the equal rights movement for literal and figurative language: A parallel search and preferential choice model. In M. Shafto & P. Langley (Eds.), *CogSci'97 – Proc. of the 19th International Conference on Cognitive Science., Stanford, CA, Aug 1997*, pp. 289–294. Mahwah, NJ: L. Erlbaum.
- Hahn, Udo, Peter Neuhaus & Norbert Bröker (1997). Message-passing protocols for real-world parsing: an object-oriented model and its preliminary evaluation. In *IWPT'97 – Proc. of the 5th International Workshop on Parsing Technologies. Boston, MA, 17–20 Sep, 1997*, pp. 101–112. Cambridge, MA: MIT Press.
- Hahn, Udo & Martin Romacker (1999a). Incrementality and locality of language comprehension: The pivotal role of semantic interpretation schemata. In *CogSci'99 – Proc.*

- of the 21st International Conference on Cognitive Science; Vancouver, Canada, Aug 1999*, pp. 202–207. Mahwah, NJ: L. Erlbaum.
- Hahn, Udo & Martin Romacker (1999b). SYNDiKATE – a generator for text knowledge bases. In T. Bench-Capon, G. Soda & A. Min Tjoa (Eds.), *DEXA '99 - Proc. 10th Intl. Conf. Database and Expert Systems Applications*, pp. 135–145. Berlin : Springer. (LN-CS, Vol. 1677).
- Hahn, Udo, Martin Romacker & Stefan Schulz (1999a). Discourse structures in medical reports - watch out! the generation of referentially coherent and valid text knowledge bases in the medSynDiKATe system,. *International Journal of Medical Informatics*, 53(1):1–22.
- Hahn, Udo & Klemens Schnattinger (1997). A qualitative growth model for real-world text knowledge bases. In *RIAO'97 - Proc. of the 1997 Conference of the Recherche d'Informations Assistee par Ordinateur*, pp. 578–597. Paris: CID.
- Hahn, Udo & Klemens Schnattinger (1998). Towards text knowledge engineering. In *AAAI'98/IAAI'98 – Proc. of the 15th National Conference on Artificial Intelligence & 10th Conf. on Innovative Applications of Artificial Intelligence. Madison, Wisconsin, 26–30 Jul 1998*, pp. 524–531. Menlo Park, CA; Cambridge, MA: AAAI Press; MIT Press.
- Hahn, Udo, Stefan Schulz & Martin Romacker (1999b). Part-whole reasoning: a case study in medical ontology engineering. *IEEE Intelligent Systems & their applications*, 14(5):59–67.
- Hahn, Udo, Stefan Schulz & Martin Romacker (1999c). Partonomic reasoning as taxonomic reasoning in medicine. In *AAAI'99/IAAI'99 – Proc. of the 16th National Conference on Artificial Intelligence & 11th Innovative Applications of Artificial Intelligence Conference. Orlando, FL, 18–22 Jul, 1999*, pp. 271–276. Menlo Park, CA; Cambridge, MA: AAAI Press & MIT Press.
- Hahn, Udo & Michael Strube (1997). Centering in-the-large: computing referential discourse segments. In *ACL'97/EACL'97 – Proc. of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter*

- of the Association for Computational Linguistics; Madrid, Spain, 7–12 Jul 1997*, pp. 104–111. San Francisco, CA: M, Kaufmann.
- Hahn, Udo, Michael Strube & Katja Markert (1996b). Bridging textual ellipses. In *COLING-96 - Proc. 16th International Conference on Computational Linguistics; Copenhagen, Denmark, 5–9 Aug 1996*, pp. 496–501.
- Haimowitz, I., R. Patil & P. Szolovits (1988). Representing medical knowledge in a terminological language is difficult. In R. A. Greenes (Ed.), *SCAMC'88 – Proc. of the 12th Annual Symposium on Computer Applications in Medical Care*, pp. 101–105. New York: IEEE Press.
- Hajicova, Eva (1987). Linguistic meaning as related to syntax and to semantic interpretation. In Makoto Nagao (Ed.), *Language and Artificial Intelligence. Proc. of an International Symposium on Language and Artificial Intelligence; Kyoto, Japan, 16–21 Mar 1986*, pp. 327–351. Amsterdam: North-Holland.
- Hartung, Joachim, Bärbel Elpelt & Karl-Heinz Klösener (1991). *Statistik : Lehr- und Handbuch der angewandten Statistik* (8th ed.). München : Oldenbourg.
- Heckl, Reiner W. (1990). *Der Arztbrief. Eine Anleitung zum klinischen Denken* (2nd ed.). Stuttgart : Thieme.
- Heinsohn, Jochen, Daniel Kudenko, Bernhard Nebel & Hans-Jürgen Profitlich (1994). An empirical analysis of terminological representation systems. *Artificial Intelligence*, 68(2):367–397.
- Henschel, Renate & John Bateman (1994). The merged upper model: a linguistic ontology for German and English. In *COLING'94 – Proc. of the 15th International Conference on Computational Linguistics; Kyoto, Japan, 5–9 Aug 1994*, pp. 803–809.
- Hersh, William R. (1996). *Information Retrieval. A Health Care Perspective*. New York: Springer.
- Hersh, William R., Emily M. Campbell, David A. Evans & Nicholas D. Brownlow (1996). Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. In James J. Cimino (Ed.), *AMIA '96 - Proc. 20th AMIA*

- Annual Fall Symposium (formerly SCAMC); Washington, D.C., 26–30 Oct 1996*, pp. 159–163. Philadelphia, PA: Hanley & Belfus.
- Hirst, Graeme (1988). Semantic interpretation and ambiguity. *Artificial Intelligence*, 34(2):131–177.
- Hirst, Graeme (1997). Context as a spurious concept. In *Proc. of the AAAI Fall Symposium on Context in Knowledge Representation and Natural Language. Cambridge, MA, 8 Nov 1997*, pp. 1–19.
- Hobbs, Jerry R., Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel & Mabry Tyson (1996). FASTUS: a cascaded finite-state transducer for extracting information from natural-language text. In Emmanuel Roche & Yves Schabes (Eds.), *Finite State Devices for Natural Language Processing*. Cambridge, MA: MIT Press.
- Hobbs, Jerry R., Douglas E. Appelt, John Bear, Mabry Tyson & David Magerman (1992). Robust processing of real-world natural-language texts. In Paul S. Jacobs (Ed.), *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pp. 13–33. Hillsdale, NJ : L. Erlbaum.
- Hobbs, Jerry R. & John Bear (1990). Two principles of parse preference. In *COLING'90 – Papers Presented at the 13th International Conference on Computational Linguistics on the Occasion of the 25th Anniversary of COLING & the 350th Anniversary of Helsinki University. Helsinki, Finland, 1990*, Vol. 3, pp. 162–167.
- Hobbs, Jerry R., Mark Stickel, Douglas E. Appelt & Paul Martin (1993). Interpretation as abduction. *Artificial Intelligence*, 63:69–142.
- Hollunder, Bernhard, Werner Nutt & Manfred Schmidt-Schauß (1990). Subsumption algorithms for concept description languages. In *ECAI'90 – Proc. of the 9th European Conference on Artificial Intelligence*, pp. 348–353. London: Pitman.
- Horrocks, Ian R. (1995). *A Comparison of Two Terminological Knowledge Representation Systems*, (Master's thesis). Univ. of Manchester.
- Horrocks, Ian R. & Peter F. Patel-Schneider (1998). DL system comparison. In Enrico Franconi, Giuseppe De Giacomo, Robert M. MacGregor, Werner Nutt & Christopher A.

- Welty (Eds.), *DL'98 – Proc. of the 1998 International Workshop on Description Logics; Trento, Italy*, pp. 55–57.
- Horrocks, Ian R., Alan L. Rector & C. A. Goble (1996). A description logic based schema for the classification of medical data. In Franz Baader, Martin Buchheit, Manfred Jeusfeld & Werner Nutt (Eds.), *KRDB'96 – Proc. of the 3rd Workshop Knowledge Representation Meets Databases*, pp. 24–28.
- Horrocks, Ian R. & Ulrike Sattler (1998a). *A Description Logic with Transitive and Converse Roles and Role Hierarchies*. Technical Report 98-05: LuFg Theoretical Computer Science, RWTH Aachen.
- Horrocks, Ian R. & Ulrike Sattler (1998b). A description logic with transitive and inverse roles and role hierarchies. In Enrico Franconi, Giuseppe De Giacomo, Robert M. MacGregor, Werner Nutt & Christopher A. Welty (Eds.), *DL'98 – Proceedings of the 1998 International Workshop on Description Logics; Trento, Italy*, pp. 72–81.
- Hripcsak, George (1994). Writing Arden syntax medical logic modules. *Computers in Biology and Medicine*, 24(5):331–363.
- Hripcsak, George, Carol Friedman, Philip O. Alderson, William DuMouchel, Stephen B. Johnson & Paul D. Clayton (1995). Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of Internal Medicine*, 122(9):681–688.
- Hull, Richard D. & Fernando Gomez (1996). Semantic interpretation of nominalizations. In *AAAI'96 – Proc. of the 13th National Conference on Artificial Intelligence; Portland, Oregon, 1996*, pp. 1062–68.
- Humphreys, Betsy L., Alexa T. McCray & May L. Cheh (1998). Evaluating the coverage of controlled health data terminologies: Report on the results of the NLM/AHCPR large scale vocabulary test. *Journal of the American Medical Informatics Association*, 4(6):484–500.
- Hunaying, Gu, Perl Yehoshua, James Geller, Michael Hanalper, Li-Min. Liu & James J. Cimino (2000). Representing the UMLS as an object-oriented database: Modeling issues and advantages. *Journal of the American Medical Informatics Association*, 7(1):66–80.

- ICD-10 (1992). *International Statistical Classification of Diseases and Health Related Problems. 10th Revision*. Geneva: WHO.
- Ingenerf, Josef (1996). On the relationship between description logics and conceptual graphs - with some reference to the medical domain. In H.-H. Bock & W. Polasek (Eds.), *Proc. of the 19th Annual Meeting of the GfKI; Basel, Mar 1996*, pp. 355–369.
- Ingenerf, Josef & T. Diedrich (1997). Notwendigkeit und Funktionalität eines Terminologieservers in der Medizin. *Künstliche Intelligenz*, 11(3):6–14.
- Jacobs, Paul S. (1991). Integrating language and meaning in structured inheritance networks. In John F. Sowa (Ed.), *Principles of Semantic Networks. Explorations in the Representation of Knowledge*, pp. 527–542. San Mateo, CA: M. Kaufmann.
- Jacobs, Paul S. & Lisa F. Rau (1990). SCISOR: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97.
- Jain, Nilesh L. & Carol Friedman (1997). Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In R. Masys (Ed.), *AMIA'97 - Proc. of the 1997 AMIA Annual Fall Symposium (formerly SCAMC); Nashville, TN, 25-29 Oct, 1997*, pp. 829–833. Philadelphia, PA: Hanley & Belfus.
- Jain, Nilesh L., Charles A. Knirsch, Carol Friedman & George Hripcsak (1996). Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. In James J. Cimino (Ed.), *AMIA'96 - Proc. 20th AMIA Annual Fall Symposium (formerly SCAMC); Washington, D.C., 26-30 Oct, 1996*, pp. 542–546. Philadelphia, PA: Hanley & Belfus.
- Johnson, Stephen B. & Carol Friedman (1996). Integrating data from natural language processing into a clinical information system. In James J. Cimino (Ed.), *AMIA'96 - Proc. of the 20th AMIA Annual Fall Symposium (formerly SCAMC); Washington, D.C., 26-30 Oct 1996*, pp. 537–541. Philadelphia, PA: Hanley & Belfus.
- Kaczmarek, Thomas S., Raymond Bates & Gabriel Robins (1986). Recent developments in NIKL. In *AAAI'86 - Proc. of the 5th National Conference on Artificial Intelligence; Philadelphia, PA, 11-15 Aug, 1986*, pp. 978–987. Los Altos, CA: M. Kaufmann.

- Kaufmann, I. (1995). Direktionale Präpositionen. In Christoph Habel, Michael Herweg & Klaus Rehkämper (Eds.), *Raumkonzepte in Verstehensprozessen*, pp. 128–149. Tübingen: M. Niemeyer.
- Klar, Rüdiger, Albrecht Zaiß, A. Timmermann & U. Schrader (1991). The information system of the Freiburger University Hospital. In K. Adlassnig & et al. (Eds.), *MIE'91 - Proc. of the 8th Conference of the Medical Informatics Europe; Vienna, Austria, Aug 1991.*, pp. 46–50. Berlin: Springer.
- Klenner, Manfred & Udo Hahn (1994). Concept versioning: a methodology for tracking evolutionary concept drift in dynamic concept systems. In Anthony G. Cohn (Ed.), *ECAI'94 - Proc. 11th European Conference on Artificial Intelligence. Amsterdam, The Netherlands, 8–12 Aug 1994*, pp. 473–477. Chichester: J. Wiley.
- Knight, Kevin & Steve R. Luk (1994). Building a large-scale knowledge base for machine translation. In *AAAI'94 - Proc. of the 12th National Conference on Artificial Intelligence*, pp. 773–778. Menlo Park, CA: AAAI Press.
- Lenat, Douglas B. (1995). CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):32–38.
- Lenat, Douglas B. & R. V. Guha (1990). *Building Large Knowledge-Based Systems. Representation and Inference in the CYC Project*. Reading, MA: Addison-Wesley.
- Levesque, Hector J. & Ronald J. Brachman (1987). Expressiveness and tractability in knowledge representation and reasoning. *Computational Intelligence*, 3(1):78–93.
- Lombardo, Vincenzo, Leonardo Lesmo, Luca Ferraris & Crispino Seidenari (1998). Incremental interpretation and lexicalized grammar. In Morton A. Gernsbacher & S.J. Derry (Eds.), *CogSci'98 - Proc. of the 20th Annual Conference of the Cognitive Science Society; Madison, 1–4 Aug, 1998*, pp. 621–626. Mahwah, NJ; London: Lawrence Erlbaum.
- Lyman, Margaret, Naomi Sager, Led Tick, Ngo T. Nhan, Francois Borst & Jean-Raoul Scherrer (1991). The application of natural-language processing to healthcare quality assessment. *Medical Decision Making*, 11(4 Suppl.):65–68.

- MacGregor, Robert M. (1993). Representing reified relations in LOOM. *Journal of Experimental and Theoretical Artificial Intelligence*, 5:179–183.
- MacGregor, Robert M. (1994). A description classifier for the predicate calculus. In *AAAI'94 – Proc. of the 12th National Conference on Artificial Intelligence*, pp. 213–220. Menlo Park, CA: AAAI Press.
- MacGregor, Robert M. & David Brill (1992). Recognition algorithms for the LOOM classifier. In *AAAI'92 – Proc. 10th National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.
- MacGregor, Robert M. & John Yen (1989). *LOOM: integrating multiple AI programming paradigms*. Technical Report: ISX Corporation, Technical Papers for LOOM Release 1.4.0, Document Version 1.4-1.
- Mann, William C. & Sandra A. Thompson (1988). Rhetorical Structure Theory: toward a functional theory of text organization. *Text*, 8(3):243–281.
- Markert, Katja (1999). *Metonymien – eine computerlinguistische Analyse*. Sankt Augustin: Infix.
- Markert, Katja & Udo Hahn (1997). On the interaction of metonymies and anaphora. In *IJCAI'97 – Proc. of the 15th International Joint Conference on Artificial Intelligence*. Vol. 2, pp. 1010–1015. San Francisco, CA: M. Kaufmann.
- Mays, Eric, Robert Weida, Robert Dionne, Laker Meir, Brian White, Chiong Lian & Frank J.. Oles (1996). Scalable and expressive medical terminologies. In James J. Cimino (Ed.), *AMIA'96 – Proc. of the 1996 AMIA Annual Fall Symposium (formerly SCAMC); Washington, D.C., 26–30 Oct 1996*, pp. 259–263. Philadelphia, PA: Hanley & Belfus.
- McCarthy, John (1993). Notes on formalizing context. In *IJCAI-93: Proc. 13th International Joint Conference on Artificial Intelligence; Chambery, France, 28 Aug–03 Sep 1993*, pp. 555–560. San Mateo, CA: M. Kaufmann.
- McCarthy, John & Sasa Buvač (1994). *Formalizing Context*. Technical Note STAN-CS-94-13: Stanford Univ.

- McCarthy, John & Saša Buvač (1998). Formalizing Context (Expanded Notes). In A. Ali-seda, R.J. van Glabbeek & D. Westerståhl (Eds.), *Computing Natural Language*, pp. 13–50. Stanford Univ., Center for the Study of Language and Information. (CSLI Lecture Notes, Vol. 81).
- McCray, Alexa T. & S. J. Nelson (1995). The representation of meaning in the UMLS. *Methods of Information in Medicine*, 34(1-2):193–201.
- McCray, Alexa T. & A. Razi (1995). The UMLS knowledge source server. In R. A. Greenes, H. E. Peterson & D. J. Protti (Eds.), *MEDINFO'95 – Proc. of the 8th World Congress on Medical Informatics; Vancouver, Canada*, pp. 144–147. Amsterdam: North-Holland.
- McCray, Alexa T., S. Srinivasan & A. Browne (1994). Lexical methods for managing variations in biomedical terminologies. In J. Ozbolt (Ed.), *SCAMC'94 – Proc. of the 18th Annual Symposium on Computer Applications in Medical Care; Washington, D.C.*, pp. 235–239. Philadelphia, PA: Hanley & Belfus.
- Mejino, Jose L. & Cornelius Rosse (1998). The potential of the Digital Anatomist foundational model for assuring consistency in UMLS sources. In Christopher G. Chute (Ed.), *AMIA'98 – Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7–11 Nov, 1998*, pp. 825–829. Philadelphia, PA: Hanley & Belfus.
- Mejino, Jose L. & Cornelius Rosse (1999). Conceptualization of anatomical spatial entities in the Digital Anatomist foundational model. In N. M. Lorenzi (Ed.), *AMIA'99 – Proc. of the 1999 AMIA Annual Symposium (formerly SCAMC); Washington, D.C., 6–10 Nov 1999*, pp. 112–116. Philadelphia, PA: Hanley & Belfus.
- Miller, George A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Miller, George A. (1998). Nouns in WordNet. In Christiane Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 23–46. Cambridge, MA: MIT Press.
- Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–245.

- Milward, D. (1992). Dynamics, dependency grammar and incremental interpretation. In *COLING'92 – Proc. of the 15th[sic! 14th] International Conference on Computational Linguistics*, pp. 1095–1099.
- Milward, David (1994). Dynamic dependency grammar. *Linguistics and Philosophy*, 17:561–604. (Special Issue on Mathematics of Language).
- Milward, David (1995). Incremental interpretation of categorial grammar. In *EACL'95 – Proc. of the 7th Conference of the European Chapter of the Association for Computational Linguistics; Dublin, Ireland, March 27-31*, pp. 119–126.
- Minsky, Marvin (1975). A framework for representing knowledge. In Patrick H. Winston (Ed.), *The Psychology of Computer Vision*, pp. 211–277. New York: McGraw-Hill.
- Montague, Richard (1974). English as a formal language. In Richmond H. Thomason (Ed.), *Formal Philosophy. Selected Papers of Richard Montague*, pp. 188–221. New Haven, London: Yale UP.
- Moore, Robert C. (1989). Unification-based semantic interpretation. In *ACL'89 – Proc. of the 27th Annual Meeting of the Association for Computational Linguistics; Vancouver, B.C., Canada, 26-29 June 1989*, pp. 33–41.
- MUC-7 (1998). *MUC-7 – Proc. of the 7th Message Understanding Conference*. <http://www.muc.saic.com>. San Diego, CA: Saic.
- Nazarenko, Adeline, Pierre Zweigenbaum, Jacques Bouaud & Benoit Habert (1997). Corpus-based identification and refinement of semantic classes. In R. Masys (Ed.), *AMIA'97 - Proc. of the 1997 AMIA Annual Fall Symposium (formerly SCAMC); Nashville, TN, 25-29 Oct 1997*, pp. 585–589. Philadelphia, PA: Hanley & Belfus.
- Neal, P., L. Shapiro & C. Rosse (1998). The Digital Anatomist structural abstraction: a scheme for the spatial description of anatomical entities. In Christopher G. Chute (Ed.), *AMIA'98 - Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7-11 Nov 1998*, pp. 423–427. Philadelphia, PA: Hanley & Belfus.
- Nebel, Bernhard (1990a). *Reasoning and revision in hybrid representation systems*. Berlin Springer.

- Nebel, Bernhard (1990b). Terminological logic is inherently intractable. *Artificial Intelligence*, 43:235–249.
- Nebel, Bernhard (1991). Terminological cycles: Semantics and computational properties. In John F. Sowa (Ed.), *Principles of Semantic Networks*, pp. 331–361. Los Altos, CA: M. Kaufmann.
- Neches, Robert, Richard Fikes, Tim Finin, Thomas Gruber, Ramesh Patil, Ted Senator & William R. Swartout (1991). Enabling technology for knowledge sharing. *AI Magazine*, 12(3):36–56.
- Neuhaus, Peter (1999). *Nebenläufiges Parsing - Ein lexikalisches verteiltes Verfahren zur performanzgrammatischen Analyse beim Textverstehen*. Sankt Augustin: Infix.
- Neuhaus, Peter & Udo Hahn (1996a). Restricted parallelism in object-oriented lexical parsing. In *COLING'96 – Proc. of the 16th International Conference on Computational Linguistics; Copenhagen, Denmark, 5–9 Aug 1996*, pp. 502–507.
- Neuhaus, Peter & Udo Hahn (1996b). Trading off completeness for efficiency: The PARSETALK performance grammar approach to real-world text parsing. In *FLAIRS'96 – Proc. of the 9th Florida Artificial Intelligence Research Symposium*, pp. 60–65.
- Nirenburg, Sergei & Victor Raskin (1987). The subworld concept lexicon and the lexicon management system. *Computational Linguistics*, 13(3-4):276–89.
- NLM (1997). *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.
- NLM (1998). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.
- O'Neil, M., C. Payne & J. Read (1995). Read Codes Version 3: A user led terminology. *Methods of Information in Medicine*, 34(1-2):187–192.
- Palmer, Martha S. (1990). Customizing verb definitions for specific semantic domains. *Machine Translation*, 5(1):5–30.
- Palmer, Martha S., Rebecca J. Passonneau, Carl Weir & Tim Finin (1993). The KERNEL text understanding system. *Artificial Intelligence*, 63(1-2):17–68.

- Paredes-Frigolett, Harold & Gerhard Strube (1996). Integrating world knowledge with cognitive parsing. In Garrison W. Cottrell (Ed.), *CogSci'96 – Proc. of the 18th Annual Conference of the Cognitive Science Society; San Diego, CA, 12–15 Jul 1996*, pp. 92–97. Mahwah, NJ: L.Erlbaum.
- Pereira, Fernando (1990). Categorical semantics and scoping. *Computational Linguistics*, 16(1):1–10.
- Pereira, Fernando C. N. & Martha E. Pollack (1991). Incremental interpretation. *Artificial Intelligence*, 50(1):37–82.
- Pisanelli, D. M., Aldo Gangemi & Geri Steve (1998). An ontological analysis of the UMLS metathesaurus. In Christopher G. Chute (Ed.), *AMIA '98 - Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7–11 Nov 1998*, pp. 810–814. Philadelphia, PA: Hanley & Belfus.
- Quantz, J. Joachim (1992). How to fit generalized quantifiers into terminological logics. In *ECAI'92 – Proc. of the 10th European Conference on Artificial Intelligence; Vienna, Austria*, pp. 473–478.
- Quantz, J. Joachim, Manfred Gehrke, Uwe Küssner & Birte Schmitz (1994). *The VerbMobil Domain Model*. KIT Report 122: Technische Universität Berlin.
- Rassinoux, Anne-Marie, R. A. Miller, Robert H. Baud & Jean-Raoul Scherrer (1998). Modelling concepts in medicine for medical language understanding. *Methods of Information in Medicine*, 37(4–5):361–372.
- Rau, Lisa F., Paul S. Jacobs & Uri Zernik (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing and Management*, 25(4):419–428.
- Rauh, Gisa (1991). *Approaches to Prepositions*. Tübingen: G. Narr.
- Rector, Alan L., Sean Bechhofer, C. A. Goble, Ian R. Horrocks, W. Anthony Nowlan & W. Danny Solomon (1997). The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, 9:139–171.

- Rector, Alan L., Andrzej Glowinski, W. Anthony Nowlan & A. Rossi-Mori (1995a). Medical-concept models and medical records: an approach based on GALEN and PEN & PAD. *Journal of the American Medical Informatics Association*, 2(1):19–35.
- Rector, Alan L. & Ian R. Horrocks (1997). Experience building a large, re-usable medical ontology using a description logic with transitivity and concept inclusions. In *Proc. of the AAAI Spring Symposium on Ontological Engineering*. Menlo Park, CA: AAAI Press.
- Rector, Alan L. & W. Anthony Nowlan (1993). The GALEN project. *Computer Methods & Programs in Biomedicine*, 45(1-2):75–78.
- Rector, Alan L., W. Anthony Nowlan & Andrzej Glowinski (1994). Goals for concept representation in the GALEN project. In C. Safran (Ed.), *SCAMC'93 – Proc. of the 17th Annual Symposium on Computer Applications in Medical Care*, pp. 414–418. New York: McGraw-Hill.
- Rector, Alan L., W. Danny Solomon, W. Anthony Nowlan, T. W. Rush, P. E. Zanstra & W. M. Claassen (1995b). A terminology server for medical language and medical information systems. *Methods of Information in Medicine*, 34(1):147–157.
- Reddig, Carola (1988). Wo trifft treffen Treffen? Zur semantischen Repräsentation nominalisierter Verben. In Harald Trost (Ed.), *4. Österreichische Artificial-Intelligence-Tagung; Wien, Austria, 29-31 Aug 1988*, pp. 43–53. Berlin Springer.
- Reimer, Ulrich (1991). *Einführung in die Wissensrepräsentation*. Stuttgart: B.G. Teubner.
- Rentier, Gerrit (1994). Dutch cross serial dependencies in HPSG. In *COLING'94 – Proc. of the 15th International Conference on Computational Linguistics; Kyoto, Japan, 5–9 Aug 1994*, pp. 818–822.
- Rogers, Jeremy E., Colin Price, Alan L. Rector, W. D. Solomin & N. Smeijko (1998). Validating clinical terminology structures: Integration and cross-validation of Read Thesaurus and GALEN. In Christopher G. Chute (Ed.), *AMIA '98 - Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7–11 Nov 1998*, pp. 845–849. Philadelphia, PA: Hanley & Belfus.

- Romacker, Martin, Katja Markert & Udo Hahn (1999). Lean semantic interpretation. In *IJCAI'99 - Proc. of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 31 Jul–06 Aug, 1999*, pp. 868–875.
- Rosse, Cornelius, Jose L. Mejino, Bharath R. Modayur, Rex Jakobovits, Kevin P. Hinshaw & James F. Brinkley (1998a). Motivation and organizational principles for anatomical knowledge representation. *Journal of the American Medical Informatics Association*, 5(1):17–40.
- Rosse, Cornelius, L.G. Shapiro & James F. Brinkley (1998b). The Digital Anatomist foundational model: Principles for defining and structuring its concept domain. In Christopher G. Chute (Ed.), *AMIA'98 - Proceedings of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7–11 Nov 1998*, pp. 820–824. Philadelphia, PA: Hanley & Belfus.
- Sachs, Lothar (1984). *Angewandte Statistik: Anwendung statistischer Methoden* (6th ed.). New York: Springer.
- Sager, Naomi, Carol Friedman & Margaret Lyman (1987). *Medical Language Processing. Computer Management of Narrative Text*. Reading, MA: Addison-Wesley.
- Sager, Naomi, Margaret Lyman, Christine Bucknall, Ngo T. Nhan & Led J. Tick (1994). Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160.
- Sager, Naomi, Margaret Lyman, Ngo T. Nhan & Led J. Tick (1995). Medical language processing: Application to patient data representation and automatic encoding. *Methods of Information in Medicine*, 34(1-2):140–146.
- Sager, Naomi, Margaret Lyman, Ngo T. Nhan, Led J. Tick, Francois Borst & Jean-Raoul Scherrer (1992). Clinical knowledge bases from natural language patient documents. In *MEDINFO'92 - Proc. of the 7th World Congress in Medical Informatics*, pp. 1375–1380.
- Salton, Gerard, James Allan & Chris Buckley (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108.

- Sattler, Ulrike (1995). A concept language for an engineering application with part-whole relations. In Alexander Borgida, Maurizio Lenzerini, Daniele Nardi & Bernhard Nebel (Eds.), *DL'95 – Proc. of the 1995 International Workshop on Description Logics; Rome, Italy*, pp. 119–123.
- Sattler, Ulrike (1996). A concept language extended with different kinds of transitive roles. In *KI'96 – Proc. of the 20th Annual German Conference on Artificial Intelligence*, pp. 333–345. Berlin: Springer.
- Schacht, Susanne (1999). *Eine aktorenbasierte Spezifikationssprache für lexikalisch verteilte Grammatiken und Parser*, (Ph.D. thesis). Univ. Erlangen.
- Schacht, Susanne & Udo Hahn (1997). A denotational semantics for joining description logics and object-oriented programming. In *SCAI'97 – Proc. of the 6th Scandinavian Conference on Artificial Intelligence; Helsinki, Finland, Aug 1997*, pp. 101–112. Amsterdam: IOS Press.
- Schacht, Susanne, Udo Hahn & Norbert Bröker (1994). Concurrent lexicalized dependency parsing: a behavioral view on PARSETALK events. In *COLING'94 – Proc. 15th International Conference on Computational Linguistics; Kyoto, Japan, 5–9 Aug 1994*, pp. 489–493.
- Schank, Roger C. (1975). *Conceptual Information Processing*. Amsterdam: Elsevier.
- Schauer, Holger (2000). From elementary discourse units to complex ones. In *Proceedings of SigDial2000 — the 1st Workshop of the Special Interest Group on Discourse and Dialogue of the Association for Computational Linguistics*, Association for Computational Linguistics. in conjunction with ACL2000, Hong Kong, China.
- Schauer, Holger & Udo Hahn (2000). Phrases as carriers of coherence relations. In *Cog-Sci'2000 – Proc. of the 22nd International Conference on Cognitive Science; Philadelphia, PA, Aug 2000*. (to appear). Mahwah, NJ: Lawrence Erlbaum.
- Schild, Klaus (1993). Combining terminological logic with tense logic. In *EPIA'93 – Proc. 6th Portuguese Conference on Artificial Intelligence*.
- Schmidt-Schauß, Manfred & Gert Smolka (1991). Attributive concept descriptions with complements. *Artificial Intelligence*, 48(1):1–26.

- Schmiedel, Albrecht (1990). A temporal terminologic logic. In *AAAI'90 – Proc. of the 8th National Conference on Artificial Intelligence; Cambridge, MA*, pp. 640–645. Menlo Park, CA: AAAI Press.
- Schmolze, James G. & W. S. Mark (1991). The NIKL experience. *Computational Intelligence*, 6:48–69.
- Schnattinger, Klemens (1998). *AQUA: Ein qualitativer, terminologischer Ansatz zur Generierung, Bewertung und Selektion von Lernhypothesen im Textverstehenssystem SYN-DIKATE*, (Ph.D. thesis). Universität Freiburg.
- Schnattinger, Klemens & Udo Hahn (1998). Quality-based learning. In *ECAI'98 – Proc. of the 13th European Conference on Artificial Intelligence; Brighton, GB, 23-28 Aug, 1998*, pp. 160–164. Chichester: J. Wiley.
- Schulz, Erich B., Colin Price & Philip J. B. Brown (1997). Symbolic anatomic knowledge representation in the Read Codes Version 3: structure and application. *Journal of the American Medical Informatics Association*, 4(1):38–48.
- Schulz, Stefan & Udo Hahn (2000). Knowledge engineering by large-scale knowledge reuse: experience from the medical domain. In Anthony G. Cohn, Fausto Giunchiglia & Bart Selman (Eds.), *KR'2000 – Proc. of the 7th International Conference on Principles of Knowledge Representation and Reasoning; Breckenridge, CO, 12-15 April 2000*, pp. 601–610. San Francisco, CA: M. Kaufmann.
- Schulz, Stefan, Udo Hahn & Martin Romacker (2000). Modelling anatomical spatial relations with description logics. In *AMIA'00 – Proc. of the 2000 AMIA Annual Fall Symposium (formerly SCAMC); Los Angeles, CA, 4-8 Nov 2000*. (to appear). Philadelphia, PA: Hanley & Belfus.
- Schulz, Stefan, Martin Romacker, Giovanni Faggioli & Udo Hahn (1999). From knowledge import to knowledge finishing: automatic acquisition and semi-automatic refinement of medical knowledge. In Brian R. Gaines, R. Kremer & Mark A. Musen (Eds.), *KAW'99 – Proc. of the 12th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop; Banff, Alberta, Canada, 16-21 Oct 1999*, pp. 7–8–1 – 7–8–12.

- Schulz, Stefan, Martin Romacker & Udo Hahn (1998). Part-whole reasoning in medical ontologies revisited – introducing SEP triplets into classification-based description logics. In Christopher G. Chute (Ed.), *AMIA '98 – Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7-11 Nov 1998*, pp. 830–834. Philadelphia, PA: Hanley & Belfus.
- Sekiguchi, Tsugio (1994). *Deutsche Präpositionen. Studien zu ihrer Bedeutungsform*. Tübingen: M. Niemeyer.
- Shapiro, Stuart C. (1979). The SNePS semantic network processing system. In N. Findler (Ed.), *Associative Networks: Representation and Use of Knowledge by Computers*, pp. 179–203. New York: Academic Press.
- Shapiro, Stuart C. & William J. Rapaport (1992). The SNePS family. *Computers & Mathematics with Applications*, 23(2-5):243–275.
- Shieber, Stuart C. (1985). Evidence against context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343.
- Shieber, Stuart C. (1986). *An introduction to unification-based approaches to grammar*. Stanford, CA: Stanford UP.
- Siegel, Sidney & John N. Castellan (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.
- Smart, J. & M. Roux (1995). A model for medical knowledge representation to the analysis of descriptive pathology reports. *Methods of Information in Medicine*, 34(4):352–360.
- Sneiderman, Charles A., Thomas C. Rindflesch & Alan R. Aronson (1996). Finding the findings: Identification of findings in medical literature using restricted natural language processing. In James J. Cimino (Ed.), *AMIA '96 – Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC); Washington, D.C., 26-30 Oct 1996*, pp. 239–243. Philadelphia, PA: Hanley & Belfus.
- Sneiderman, Charles A., Thomas C. Rindflesch & Carol A. Bean (1998). Identification of anatomical terminology in medical text. In Christopher G. Chute (Ed.), *AMIA '98 – Proc. of the 1998 AMIA Annual Fall Symposium; Orlando, FL, 7-11 Nov, 1998*, pp. 428–432. Philadelphia, PA: Hanley & Belfus.

- Sondheimer, Norman K., Ralph M. Weischedel & Robert J. Bobrow (1984). Semantic interpretation using KL-ONE. In *COLING'84/ACL'84 – Proc. of the 10th International Conference on Computational Linguistics & 22nd Annual Meeting of the Association for Computational Linguistics; Stanford, CA, 2–6 Jul, 1984*, pp. 101–107.
- Sowa, John F.. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.
- Sowa, John F. (1991). Toward the expressive power of natural language. In John F. Sowa (Ed.), *Principles of Semantic Networks*, pp. 157–189. Los Altos: M. Kaufmann.
- Spackman, Kent A. & K.E. Campbell (1998). Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies. In Christopher G. Chute (Ed.), *AMIA'98 – Proc. of the 1998 AMIA Annual Fall Symposium (formerly SCAMC); Orlando, FL, 7–11 Nov, 1998*, pp. 740–744. Philadelphia, PA: Hanley & Belfus.
- Spyns, Peter (1996). Natural language processing in medicine. *Methods of Information in Medicine*, 35(4-5):285–301.
- Strube, Michael (1997). *Funktionales Centering*, (Ph.D. thesis). Univ. Freiburg.
- Strube, Michael & Udo Hahn (1995a). PARSETALK about sentence- and text-level anaphora. In *EACL'95 – Proc. 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 237–244. Dublin: Association for Computational Linguistics.
- Strube, Michael & Udo Hahn (1995b). PARSETALK about sentence- and text-level anaphora. In *Proc. of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 237–244.
- Strube, Michael & Udo Hahn (1996). Functional centering. In *ACL'96 - Proc. 34th Annual Meeting of the Association for Computational Linguistics; Santa Cruz, CA, 23–28 Jun 1996*, pp. 270–277. San Francisco, CA: M. Kaufmann.
- Strube, Michael & Udo Hahn (1999). Functional centering: grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

- Sundheim, Beth (Ed.) (1991). *MUC-3 – Proc. of the 3th Message Understanding Conference*. San Mateo, CA: M. Kaufmann.
- Sundheim, Beth (Ed.) (1992). *MUC-4 – Proc. of the 4th Message Understanding Conference*. San Mateo, CA: M. Kaufmann.
- Sundheim, Beth (Ed.) (1994). *MUC-5 – Proc. of the 5th Message Understanding Conference*. San Mateo, CA: M. Kaufmann.
- Sundheim, Beth (Ed.) (1995). *MUC-6 – Proc. of the 6th Message Understanding Conference; Columbia, Maryland, November 6-8, 1995*. San Mateo, CA: M. Kaufmann.
- van Eijck, Jan & Robert C. Moore (1992). Semantic rules for English. In Hiyan Alshawi (Ed.), *The Core Language Engine*, pp. 83–115. Cambridge, MA: MIT Press.
- Volot, F., Pierre Zweigenbaum, Bruno Bachimont, M. Ben Said, Jacques Bouaud, M. Fieschi & Jean-Francois Boisvieux (1994). Structuration and acquisition of medical knowledge: using UMLS in the conceptual graph formalism. In C. Safran (Ed.), *SCAMC'93 – Proc. of the 17th Annual Symposium on Computer Applications in Medical Care*, pp. 710–714. New York: McGraw-Hill.
- Wedekind, J. & R. Kaplan (1993). [Type-driven semantic interpretation of f-structures $\langle\langle J, W \rangle, \langle R, K \rangle\rangle$]. In *EACL'93 – Proc. of the 6th Conference of the European Chapter of the ACL; Utrecht, The Netherlands, 21–23 Apr 1993*, pp. 404–411.
- Winston, Morton, Roger Chaffin & Douglas J. Herrmann (1987). A taxonomy of part-whole relationships. *Cognitive Science*, 11:417–444.
- Wolff, Suzanne (1984). The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, 23(4):195–203.
- Woods, William A. & James G. Schmolze (1992). The KL-ONE family. *Computers & Mathematics with Applications*, 23(2-5):133–177.
- Yen, John, Robert Neches & Robert MacGregor (1991). CLASP: integrating term subsumption systems and production systems. *IEEE Transactions on Knowledge and Data Engineering*, 3:25–32.

- Zaiß, Albrecht, H. Schäfer & Rüdiger Klar (1991). Experiences with PATSY, an information system for pathology. In K. Adlassnig & et al. (Eds.), *MIE'91 - Proc. of the 8th Conference of Medical Informatics Europe; Vienna, Austria, Aug 1991*, pp. 106–110. Berlin: Springer.
- Zaiß, Albrecht, Stefan Schulz, B. Graubner & Rüdiger Klar (1996). Conversion table between ICD-9 and ICD-10. In J. Brender, J. P. Christensen, Jean-Raoul Scherrer & P. McNair (Eds.), *MIE'96 - Proc. of 13th Conf. of Medical Informatics Europe*. Amsterdam: IOS press.
- Zarri, Gian P. (1995). Internal and external knowledge context, and their use for interpretation of natural language. In *Proc. of the IJCAI Workshop on Context in Natural Language Processing; Aug 1995*, pp. 180–188.
- Zweigenbaum, Pierre, Bruno Bachimont, Jacques Bouaud, Jean Charlet & Jean-Francois Boisvieux (1995a). Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods of Information in Medicine*, 34(1-2):15–24.
- Zweigenbaum, Pierre, Bruno Bachimont, Jacques Bouaud, Jean Charlet & Jean-Francois Boisvieux (1995b). A multi-lingual architecture for building a normalized conceptual representation from medical language. In R. M. Gardner (Ed.), *SCAMC'95 – Proc. of the 19th Annual Symposium on Computer Applications in Medical Care*, pp. 357–361. Philadelphia, PA: Hanley & Belfus.
- Zweigenbaum, Pierre, Jacques Bouaud, Bruno Bachimont, Jean Charlet & Jean-Francois Boisvieux (1997). Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. In R. Masys (Ed.), *AMIA'97 – Proc. of the 1997 AMIA Annual Fall Symposium (formerly SCAMC); Nashville, TN, 25–29 Oct 1997*, pp. 590–594. Philadelphia, PA: Hanley & Belfus.
- Zweigenbaum, Pierre & Consortium MENELAS (1994). MENELAS: An access system for medical records using natural language. *Computer Methods & Programs in Biomedicine*, 45(1-2):117–120.

