PARADICE

Parametrizable Discourse Core Engine

BMBF FKZ ITW 9403

Abschlußbericht

Dipl. Inf. Bernd Kiefer
Dr. Klaus Netter
Dr. Günter Neumann
Prof. Dr. Hans Uszkoreit

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Forschungsbereich Sprachtechnologie
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Tel.: 0681-302-5282
Fax.: 0681-302-5338

Juni 1997

Inhaltsverzeichnis

l Aufgabenstellung

1.1 Hauptziele

Die Entwicklung eines brauchbaren Kernsystems für natürliche Sprache (NL) benötigt viel Zeit. Die Systeme, die man heute für angewandte Arbeit im Bereich von NL verwendet, wurden über einen Zeitraum von acht bis zwölf Jahren entwickelt. Systeme, die gegenwärtig in konkreten Anwendungen benutzt werden, hatten Entwicklungszeiten von zwischen zehn und 25 Jahren. Mit dem Entstehen neuer Techniken für die Beschreibung sprachlicher Phänomene beginnt sich die minimale Entwicklungszeit nach und nach zu verringern - und doch sind wir von einer Methodologie für effizientes *linguistic engineering* noch weit entfernt.

Die grundlegenden Kernkomponenten eines fortgeschrittenen NL-Systems sind im Laufe der letzten drei Jahre des DFKI-Projekt DISCO entwickelt worden. Da DISCO das erste große NL-Projekt am DFKI war, wurde jede Systemkomponente von Grund auf entwickelt, wobei man die fortschrittlichsten Methoden und Techniken anwandte, die zur Verfügung standen. Es wurde eine neue Entwicklungsumgebung geschaffen, die auf Konferenzen und Workshops große Anerkennung fand, und die bereits begonnen hat, Früchte zu tragen. Die Aufgabe des Projektes PARADICE ist es, diese Arbeit fortzuführen. Ziel ist dabei, das existierende NL-Kernsystem zu einem flexiblen, parametrisierbaren Diskurssystem, das für eine große Vielfalt von Anwendungen nutzbar gemacht werden kann, weiterzuentwickeln. Das PARADICE Projekt unterscheidet sich von seinem Vorgänger darin, daß es sich auf eine effiziente und robuste Ausnutzung des Kernsystems konzentriert. Das bedeutet, daß alle Kernkomponenten des DISCO-Systems erhalten bleiben, während ihre Funktionalität, Leistung und Interaktionen in großem Umfang erweitert werden.

Der Ansatz, den man in DISCO zugrundegelegt hat, und der mit Nachdruck in PARADICE weiterverfolgt werden sollte, unterscheidet sich stark von den meisten früheren NL-Projekten. Er basiert auf einer strikten Trennung zwischen einem Kompetenzsystem auf der einen und Performanzsystemen auf der anderen Seite. Im Kompetenzsystem wird linguistisches Wissen in einem modernen logik-basierten Spezifikationsformalismus höherer Stufe beschrieben. Die Verarbeitungsalgorithmen sind vollständig und korrekt hinsichtlich der linguistischen Spezifikation. Die Spezifikation hält sich an die Standards von adäquater linguistischer Beschreibung, wie sie gegenwärtig in der theoretischen formalen Linguistik definiert wird. Sie macht keinerlei Kompromiß bei der Verarbeitungseffizienz.

Es ist ziemlich offensichtlich, daß das Kompetenzsystem die Anforderungen, wie sie von einer realistischen Anwendung gestellt werden, nicht erfüllt. Es soll vielmehr als Quelle dienen, aus der man linguistische Spezifikation für ein Performanzsystem mit Hilfe von Refokussierung, Kompilation und durch eine Änderung des Sprachumfangs ableiten kann. Der Sprachumfang kann reduziert werden, wenn die Anwendung nicht die gesamte Kapazität des Kompetenzsystems benötigt. Andernfalls können spezielle Regeln oder Lexikoneinträge, die nicht in der Standard-Kompetenzgrammatik der Sprache vorliegen, für einen speziellen Bereich hinzugefügt werden. Für eine effiziente Verarbeitung kann man die Performanzgrammatik in ein Format kompilieren, das leichter zu verarbeiten ist, so z.B. Mengen von Bäumen, große Mengen von Phrasenstrukturregeln oder sogar endliche Automaten. Lernverfahren, die auf empirischer Messung in einem Bereich basieren, können verwendet werden, um die Verarbeitung zentraler Beispiele zu beschleunigen, wobei teurere Methoden peripheren Konstruktionen vorbehalten bleiben.

In DISCO lag der Schwerpunkt darauf, ein Kompetenz-Kernsystem für die deutsche Spra-

ehe zu entwickeln. Das PARADICE-Projekt konzentriert sich darauf, dieses Kernsystem für Anwendungen nutzbar zu machen, indem Performanzmethoden für eine robuste und effiziente Verarbeitung auf der Grundlage eines einheitlichen, allgemeinen, eleganten und deklarativen Kompetenzsystems entwickelt werden, und indem die bereits vorhandenen begrenzten Diskurs- Fähigkeiten zu einem allgemeinen Modell der Dialogkompetenz ausgearbeitet werden.

Die Erfahrung hat gezeigt, daß die langen Entwicklungszeiten von NL-Systemen es unmöglich machen, die gesamte Vielfalt ihrer möglichen Anwendungen vorherzusehen. Aus diesem Grund muß man von Anfang an Flexibilität in das Kompetenz-Kernsystem einbauen. PARADICE macht sich einen Parametrisierungsansatz zu eigen, der den möglichen Anwendungsbereich des Kernsystems und seinen an die Performanz angepaßten Ableitungen maximiert. Um die nötige Erfahrung zu sammeln, wie diese Flexibilität eingesetzt werden kann, um den Entwurf bewerten zu können und schließlich um die praktische Anwendbarkeit zu zeigen, ist es nötig, Demonstrationsanwendungen zu entwickeln. Zwei Klassen von solchen Anwendungsbeispielen waren geplant, wobei eine intern in PARADICE und die andere in Abstimmung mit dem beantragten Schwesterprojekt COSMA durchgeführt werden sollte.

1.2 Anwendungsbereiche

Die Flexibilität eines Systems kann man schwerlich an einer einzigen Anwendung demonstrieren. So ist beim gegenwärtigen *state-of-the-art* wenig über ein NL-Diskurs-Kernsystem gesagt, wenn man zeigt, daß es sowohl über Dieselmotoren als auch über Terminkalender verhandeln kann: Das Beherrschen zweier Wissensdomänen sagt viel über Anwendungen und das Lexikon, jedoch wenig über die Anpassungsfähigkeit der gesamten linguistischen Fähigkeit. Ein echter Flexibilitätstest in einem System wie PARADICE besteht darin, zu zeigen, wie mit verschiedenen *Typen* von Diskurs in verschiedenen Anwendungen umgegangen wird.

Neben der linguistischen Abdeckung im engeren Sinne sollten also Anwendungsbeispiele verfolgt werden, die Fragen beantworten können wie: Ist das System anpassungsfähig an verschiedenen Kommunikationsmedien, wie z.B. der Diskurssituation bei elektronischer Post im Gegensatz zu dem gezielten Verstehen von Passagen in grösseren Texten? Kann es mit variierenden Annahmen darüber, wer die Initiative ergreift (das System oder der Benutzer), fertig werden? Versteht es die Konventionen, nach denen NL-Diskurs einmal aus langen Blöcken von sorgfältig zusammengestellter Prosa, ein anderes Mal aus kurzen Stücken von sich gegenseitig unterbrechenden Texten besteht, und kann es dieses Wissen ausnutzen?

Angesichts der oben angeführten Fragestellungen, um die Forschung zu konzentrieren und um ferner die praktische Anwendbarkeit zu zeigen, wurden zwei Anwendungsbereiche ausgewählt, von denen die erste im Schwesterprojekt COSMA ausgeführt wurde, die zweite im PARADICE-Projekt selbst realisiert wurde.

- Diskurse bei elektronischer Post in einer beschränkten Domäne, so daß eine Vielfalt von ein- und zweiseitiger Kommunikation zwischen beiden Paaren und Gruppen abgedeckt wird; und

- Gezielte Extraktion von Informationen aus Texten in verschiedenen Gegenstandsbereichen.[1]

[1]In einer allerersten Stufe der Beantragung war hier als zweite Anwendung noch ein "Informelle Mensch-Maschine-Dialog mit zwei Partnern für Datenbankzugriffe" geplant. Relativ früh hat sich jedoch gezeigt, daß durch die starke Dialog-Orientiertheit beider Anwendungen die Bandbreite der Anwendungen der Kernma-

Der erste Bereich, die Verarbeitung von E-Mail-Messages, wurde vorwiegend aus praktischen Erwägungen heraus ausgewählt:

- In vorhergehender Arbeit mit elektronischer Post im Projekt DISCO konnte wichtige praktische Erfahrung gesammelt werden;
- Zugang zu beträchtlichen Test-Korpora und die Konstruktion von hausinternen Tests sind relativ einfach;
- semi-automatische Verarbeitung von elektronischer Post wird als eine bedeutende praktische Anwendung von NL-Technologie in nächster Zukunft vorausgesehen;
- da sich die Projekte ALV und WIDAN am DFKI Standort Kaiserslautern mit partieller Interpretation von *gedruckten* oder *maschinengeschriebenen* Bürodokumenten befassen, sind die Möglichkeiten zur internen Zusammenarbeit hervorragend.

Sich mit elektronischer Post zu befassen, wird umso attraktiver, je stärker die Akzeptanz dieses Kommunikationsmediums zunimmt. In vielen Bereichen ist sie im Begriff, die Rolle der bevorzugten Form innerhalb der geschäftlichen Kommunikation zu übernehmen. Die Anwendungsbereiche gehen dabei von faktisch unrestringierter Kommunikation, wie z.B. Diskussionen über Personalentscheidungen, bis hin zu vollkommen standardisierter, inklusive Rechnungen und Versandaufträge. Während der erste Typ für derzeitige NL-Software zu schwierig ist, und man den zweiten durch spezialisierte Verarbeiter mit genauer Kenntnis des standardisierten Repräsentationssystems effizienter handhaben kann, gibt es schließlich ein breites Mittelfeld: Viele Domänen sind beschränkt genug, um die Verwendung von NL-Technlogie zu erlauben, und sind andererseits weit entfernt davon, vollkommen standardisierbar zu sein. Dialoge über Verabredungen oder Zeitpläne, Verhandlungen über Versandaufträge sowie standardisierte geschäftliche Mitteilungen mit informellen Anteilen gehören zu dieser Klasse. Eine große Vielfalt von Software-Werkzeugen wie Kalendersysteme, Verabredungs- und Versandplaner, Verarbeiter von Bestellungen etc. werden schließlich den Menschen darin unterstützen, diese Typen der Kommunikation zu verarbeiten. In all diesen Fällen werden NL-Fähigkeiten und -Möglichkeiten für das Verarbeiten, Vor-Verarbeiten, Klassifizieren, sowie dafür, elektronisch übermittelte Information zu verteilen und abzufragen, ganz wesentlich sein.

Die erste Anwendung (und Gegenstand des Schwester-Projektes COSMA) ist die Konsolidierung und Erweiterung des existierenden, auf elektronischer Post basierenden Kalenderagenten COSMA sein. Zusätzlich zu der Menge der Einstellungen von Diskurs-Parametern, die für elektronische Post mit mehreren Teilnehmern geeignet sind, verlangt ein Zuwachs in der Funktionalität einfache Experimente mit anderen Punkten des Koordinatensystems, wie z.B. Erinnerungen eines einzelnen Agenten, die keine Antwort, oder erlaubt öffentlich verbreitete Ankündigungen.

Das Treffen und Zusammenstellen von Verabredungen und die Verarbeitung von Ereignisankündigungen bilden eine angemessene und hinreichend restringierte Domäne, die sich als plausibel erwiesen hat. Es wird immer einen Bedarf für NL-Kommunikation in dieser Anwendung geben - aufgrund von Benutzern, die nicht in Besitz eines Systems sind, das

schine stark eingeschränkt worden wäre. Die Richtigkeit der Wahl der Informationsextraktion als zweite Anwendungsdomäne wurde unter anderem durch den großen Aufschwung des Gebietes in den letzten Jahren und auch das industrielle Interesse an dem aus dem PARADICE-Projekt entstandenen Prototyp bestätigt.

Verabredungen organisiert, oder die es nicht benutzen können, aufgrund fehlender Protokollstandards für verschiedene Terminmanager, ferner weil ein menschliches Überprüfen von automatisch geführten Verhandlungen in kritischen Fällen vonnöten ist, und schließlich - und das ist für die Zukunft besonders wichtig - weil es wesentlich ist, daß Menschen ein starkes Gefühl für die Verbindung mit und die Kontrolle über automatisierte Systeme behalten, denen eine zunehmend bedeutende Rolle im Organisieren ihres Lebens zukommt.

Während im COSMA-Projekt die Motivation darin besteht, eine vielversprechende Forschungsrichtung in der Anwendung selbst weiterzuverfolgen, war das Ziel der zweiten Anwendung, mit einer minimalen Abzweigung von Ressourcen von den Forschungszielen des Projektes maximale praktische Erfahrungen zu sammeln.

Anstelle der Datenbank-Applikation wurde deshalb in der zweiten Hälfte des Projektes eine unabhängige Anwendungsrichtung aufgebaut, nämlich die Extraktion und gezielte Identifikation von Informationen in elektronischen Texten, kurz *Message Extraction.*

Die Motivation für die Wahl dieser Anwendung als Komplement und Ergänzung zur COSMA-Anwendung bestand u.a. in den folgenden Punkten:

- Ein NL-Kern-System darf nicht auf die Verarbeitung von Dialogen beschränkt sein, sondern muß auch deskriptive Texte bearbeiten können, die über die Länge von Dialog-Turns hinausgehen.
- Eine Vielzahl von Informationen wird heutzutage in Form von Texten mittlerer Läenge (ca. eine Seite) angeboten. Wohingegen ein vollständiges Verstehen solcher Texte auf absehbare Zeit ausgeschlossen erscheint, ist die gezielte Identifikation und Extraktion von Informationen durchaus im Bereich des Machbaren.
- Während beim Dialog in der Regel eine tiefe Verarbeitung und ein tiefes Verstehen vorausgesetzt wird, erfordert die gezielte Extraktion von Informationen eher eine robuste Verarbeitung, die auch mit einer eher "flachen" Analyse auskommt.

Anwendung für Informationsextraktion findet man in den verschiedensten Bereichen. Bereits seit längerer Zeit finden zu diesem Anwendungsfeld in den USA die MUC-Wettbewerbe (Message Understanding Conference) statt, in denen so verschiedene Domänen wie Ausfallberichte von militärischem Material, Berichte zu Terroristenüberfäellen, Wirtschaftsmeldungen über Joint Ventures und dergleichen zum Gegenstandsbereich gewählt wurden.

Die Bandbreite der Domänen, die im PARADICE-Project behandelt werden können, sind natürlich stark eingeschränkt und ergeben sich vor allem aus dem Zugang zu geeigneten Korpora. Im wesentlichen werden zwei Gebiete behandelt, die Extraktion von Daten aus Ankündigungen zu Veranstaltungen, die vergleichsweise stark strukturiert sind, sowie die Extraktion von Informationen zu Umsatzzahlen oder Jahresabschlüssen in Wirtschaftsmeldungen.

Eine Parametrisierung des Kernsystems ergibt sich sich durch diese zweite Anwendung damit nicht nur hinsichtlich der Diskurs-Struktur, sondern auch hinsichtlich der Verarbeitungstiefe und hinsichtlich des Umfangs und der thematischen Breite der zu verarbeitenden Texte. Die verschiedenen Verarbeitungsmodi in einem System bereitzustellen und kombinieren zu können war eine der Herausforderungen des Projekts.

1.3 Wissenschaftliche Motivation

Eines der wichtigsten ungelösten Probleme von *language engineering* - vielleicht sogar das zentrale offene Problem - ist der Konflikt zwischen der neuen Klasse von Repräsentations-

formalismen höherer Stufe und dem dringenden Bedarf an effizienten und robusten Verarbeitungssystemen.

Wie bereits oben ausgeführt, war *linguistic Engineering* in der Vergangenheit übermäßig, fast untragbar teuer. Dies lag zum Teil an mangelnder Leistungsfähigkeit, hervorgerufen durch das Fehlen von geeigneten Methoden für verteiltes *engineering*, und zum Teil an der Schwierigkeit, operational konzeptualisierte (aber nicht computerorientierte) grammatische Theorien zu prozeduralen Implementierungen in Beziehung zu setzen.

Mit dem Aufkommen von in hohem Grade deklarativen unifikations-basierten Formalismen hat sich die Situation extrem verbessert; fast jedes neue Projekt im Bereich moderner NL-Verarbeitung setzt unifikations-basierte Formalismen zur grammatischen Beschreibung ein. Diese neueren Formalismen bieten eine saubere modelltheoretische Semantik und erschließen neue Möglichkeiten für effektiveres *grammar engineering.* Insbesondere die verteilte Grammatikentwicklung ist ein erreichbares Ziel geworden. Zum ersten Mal gibt es Grund zur Hoffnung, daß beträchtliche wiederverwendbare linguistische Ressourcen geschaffen werden können. Es wurden bereits mit anderen Gruppen, die unterschiedliche Kernkomponenten verwenden, Grammatiken ausgetauscht. Noch wichtiger für das Verfolgen von wissenschaftlichen Zielen ist es, daß die neuen Formalismen den Graben zwischen linguistischer Beschreibung in der formalen theoretischen Lingistik und in *linguistic engineering* überbrückt haben, und so einen direkten Transfer von Ergebnissen zwischen theoretischer Lingistik und Computerlinguistik erlauben.

Nach einem Jahrzehnt Forschung stehen für Unifikationsformalismen jedoch immer noch keine effizienten Verarbeitungsmethoden zur Verfügung. Forscher an zahlreichen Orten arbeiten an besseren Verarbeitungsalgorithmen, aber alle Projekte, die Effizienz beibehalten müssen, verwenden linguistische Beschreibungssprachen niedrigerer Stufe, die zugunsten der Geschwindigkeit gegenüber der Theorie Kompromisse eingehen. Effizienz und Robustheit müssen jedoch auch erreicht werden können, *ohne* daß die Vorteile von gut handhabbaren, anspruchsvollen und linguistisch adäquaten Formalismen geopfert werden müssen. Dieses Ziel strebt das Projekt PARADICE an.

In DISCO wurde bereits in konzertierter Weise begonnen, bessere Methoden für grammatische Kompilierung und Kontrolle zu entwickeln. PARADICE verfolgt diesen Weg weiter. Während DISCO vorwiegend darauf ausgerichtet war, den zentralen Kern eines linguistischen Kompetenzsystems aufzubauen, konzentriert sich PARADICE darauf, dieses Kernsystem ohne Kompromisse zu einer Kernmaschine mit großem Sprachumfang und mit effizienten und robusten abgeleiteten Performanzmodellen zu erweitern, um so dem klassischen Defizit bei der Performanz begegnen.

Weiterhin folgt PARADICE seinem Vorgängerprojekt in der Grundannahme, daß Diskursphänomene zum Bereich der linguistischen Kernmaschine gehören, und sie nicht etwa an ein möglicherweise - möglicherweise auch nicht - intelligentes *back end* verbannt werden. Das DISCO-Projekt befaßte sich mit dem Entwurf von linguistischen Kompetenz-Modulen, die sämtliche Ebenen linguistischen Wissens - inklusive der linguistischen Aspekte von Diskurskompetenz - auf relativ einheitliche Weise integrieren. Eine Herausforderung für das neue Projekt ist es, diesen einheitlichen Ansatz auszunutzen und weiterzuverfolgen.

Eine spezifische Verbindung zwischen dem Kernsystem und einer Anwendung wird in aller Tiefe im parallelen CosMA-Projekt erforscht. Dort wird ferner der Begriff der NL-Kernmaschine als kooperativer linguistischer Agent untersucht. PARADICE stellt die kompromißlose Basis dazu bereit.

2 Voraussetzungen zur Durchführung des Vorhabens

Die Implementierungen und theoretischen Ergebnisse, die aus den DISCO und ASL Projekten hervorgegangen sind, stellen eine solide Grundlage für die Arbeit in PARADICE dar.

Der zentrale Prozessor des *constraint-basierieu* Formalismus im DISCO-System ist der Unifikator UDINE, der volle Negation, verteilte Disjunktion und relationale *constraints* zur Verfügung stellt (zu zugrundeliegenden Ideen siehe [BEG90]) und [BS93] und damit über die meisten anderen Unifikatoren bei weitem hinaus geht. Aufbauend auf diesem Unifikator wurde der getypte Merkmalslogik-Formalismus TDL ([KS93]) entwickelt, der eine deklarative Spezifikationssprache darstellt, die über hinreichende Ausdrucksstärke für theoretisch anspruchsvolle Spezifikationen von Grammatikfragmenten verfügt. Seine Eigenschaften beinhalten multiple Typvererbung, boolesche Kombinationen von Typen und Merkmalsstrukturen, Partition und andere Formen von Unvereinbarkeits-Deklarationen, parametrisierte Templates sowie partielle oder verzögerte Typenexpansion. Wichtige theoretische Beiträge im Formalismusbereich schließen den Entscheidbarkeitsbeweis von Konsistenztests für Klauseln ein, die *functional uncertainty* beinhalten [Bac93a], [Bac93b]. Erweiterungen des Typsystems, um die Spezifikation von regulären Sprachen einzuschließen, wie sie z.B. in einer Zweistufenmorphologie benutzt werden, wurden bei [KNP93] gezeigt.

Der Formalismus wird ergänzt durch eine der umfangreichsten existierenden Umgebungen für Grammatikentwicklung. Diese beinhaltet vor allem den Merkmals-Editor FEGRAMED, der auch außerhalb des Projekts weite Verbreitung fand und in die ALEP Umgebung integriert ist, ein Standard Softwarepaket für NLP, das mit Unterstützung der Europäischen Kommission (CEC) vertrieben wird. Ein wichtiger Beitrag auf dem Gebiet des Überprüfens und Bewertens von entwickelten Grammatiken geht zurück auf eine Initiative von DISCO-Mitarbeitern, eine große Testsuite für das Deutsche zu erstellen, das DiTo-System [KDD+92]. Diese Initiative wurde von verschiedenen industriellen und akademischen Istitutionen aufgegriffen und bildete ferner die Grundlage für das TSNLP Projekt, das im Rahmen des EU-LRE Programms durchgeführt wird.

Die zentrale Verarbeitungsmaschine enthält einen parametrisierbaren bidirektionalen Chart Parser, der die deklarative Spezifikation von Kontrollstrategien erlaubt. Für die Oberflächengenerierung ist ein vom semantischen Kopf gesteuerter Ansatz implementiert worden, für den man ein Diskursgedächtniss als kontextuelle Wissensbasis nutzen kann, um die Ausgabe zu beschränken. Für die morphologische Verarbeitung wurde ein Zwei-Ebenen-Automaten-Modell mit Merkmalsbeschränkungen als Filter entwickelt [Tro90]. Eine objekt-orientierte Architektur wurde als flexible Plattform für die Systemintegration gewählt, die großes Gewicht auf die Kontrolle des Informationsflusses zwischen Modulen legt und deren Integration erheblich vereinfacht und beschleunigt [Neu93]. Theoretische Ergebnisse und prototypische Implementierungen zu effizienten Kontrollstrategien stehen zur Verfügung und sind in [Usz91] und [Bac92] beschrieben. Ein erster Prototyp für die effiziente Verarbeitung von Subsprachen wurde auf der Grundlage der *explanation based learning* Methode (EBL) realisiert.

Aufbauend auf dieser Umgebung wurde ein wesentliches Fragment der deutschen Grammatik spezifiziert, das intensiv Gebrauch macht von Ergebnissen theoretischer Arbeiten, die im Bereich von theoretischer Linguistik und Computerlinguistik verwendet werden. Der höhere Formalismus und die solide theoretische Grundlage der linguistischen Spezifikationen führen zu einem hohen Maß an Transparenz und Modularität in der Grammatik, was eine unerläßliche Voraussetzung für Wiederverwendbarkeit und verteilte Grammatikentwicklung darstellt [Net93a]. Einige der theoretisch interessanten Aspekte der Grammatik sind dokumentiert in

[Net92] und [Net93b]. Theoretische Ergebnisse über die Integration von derivationeller Morphologie und dem Lexikon im Rahmen der HPSG sind beschrieben in [KN92].

In enger Zusammenarbeit mit dem ASL-Projekt ist eine Syntax/Semantik-Schnittstelle entwickelt worden, die vollständig in die Grammatik integriert ist. Sie unterstützt die syntaktische Analyse und Disambiguierung anhand von Sorten-Beschränkungen ([Ner91a], [Ner92], [Ner91b]) ebenso wie die Auflösung von Skopus und Anaphern auf der Ebene der logischen Form. Allgemeine Werkzeuge wurden entwickelt, die Techniken zur Transformation von Programmen anwenden, um semantische Merkmalsstrukturen in aufgelöste logische Repräsentationen mit Skopusinformation als Schnittstelle für Anwendungen umzuwandeln [NLDO92], [NOD+93]. Eine Erweiterung dieser Arbeit war die Entwicklung einer allgemeinen Sprache für die deklarative Spezifikation von Transformationen, wie sie bei Schnittstellen verwendet werden ([BKN+93]). Eine Möglichkeit für die Resolution von Anaphern, die Ideen von aktuellen syntaktischen und semantischen Theorien aufnimmt, ermöglicht die Verarbeitung von Diskursen und Dialogen mit mehreren Sätzen [Kas93], aufbauend auf einer dynamischen Interpretation von logischen Repräsentationen für Diskurs ([Kas]).

Die Diskursfähigkeiten, die im DISCO-Projekt entwickelt wurden, beinhalten ein Multi-Agenten-Modell von Glaubenszuständen [SH93], n-fach verläßliche Sprechakte [HS92], einfache Mehrsatz-Sprechakte und einfache Interaktionen von mehreren Sprechakten. Die Erkennung von Sprechakten stützt sich auf detaillierte syntaktische Information ebenso wie auf inferentiellen Kontext und basiert so auf enger Integration von linguistischen Prozessen. Diese Fähigkeiten sind hinreichend, um Dialoge mit mehreren Agenten über Terminabsprachen zu handhaben.

Ferner stellt die COSMA-Anwendung, die als eine Demonstration der Anwendbarkeit des DISCO-Systems entwickelt wurde, ein plausibles Modell für die gemeinsame Verwendung einer graphischen Benutzeroberfläche und natürlicher Sprache dar, und hebt die Verarbeitung natürlicher Sprache im Kontext der Aktivität einer kooperativen Gruppe hervor, die aus autonomen Software-Agenten und menschlichen Benutzern besteht, wobei diese über spezialisierte Software-Vermittler verfügen oder nicht verfügen können.

Die Computerlinguistik-Abteilung des DFKI (Projekte DISCO und ASL) organisierte die drei bedeutendsten internationalen Workshops in Deutschland, die in direkter Beziehung stehen zu den Zielen des Projektes: den *International Workshop on Grammar Engineering 1990* [EU90], den *International Workshop on HPSG for German 1991* [NNP93] und den *International Workshop on Implemented Grammar Formalisms* im März 1993. Das DFKI leitete die *Expert Group on Linguistic Formalisms* (Vorsitzender Hans Uszkoreit) der CEC-Initiative EAGLES *(Expert Advisory Groups on Linguistic Engineering Standards),* in denen auch zwei der DFKI-Gesellschafter involviert waren.

Mitglieder des DISCO-Projektes (Backofen, Netter, Uszkoreit) sind beteiligt an der Spezifikation des Formalismus und der Entwicklungsplattform des ALEP-Systems, das von der belgischen Firma BIM, später von Cray Systems, für die CEC entwickelt wird.

3 Planung und Ablauf des Vorhabens

3.1 Planung des Vorhabens

Die Planung des Projektes hatte als Ziel ein natürlichsprachliches Kernsystem zu konsturieren, das als Grundlage für verschiedene Anwendungen dienen kann. Dieses Kernsystem ist in zwei Bereiche (die linguistische Wissensbasen und die Verarbeitung) getrennt. Die linguistische Wissensbasen umfassen Morphologie, Lexikon, Grammatik und Diskurs. Die Verarbeitungskomponenten lassen sich in Kompetenz- und Performanz-orientierte Komponenten aufteilen, wobei die ersteren die Basisfunkionalität realisieren (Scanner, morphologische, lexikalische und grammatische Analyse und Synthese) und die letzteren speziell zur effizienten und robusten Verarbeitung dienen (Explanation Based Learning - EBL, Subsprachen, Präferenzbasierte Methoden, flexible Architektur). Da das System als Kernsystem in COSMA integriert wird, umfaßt es darüberhinaus auch die Basis für Diskurs und entsprechende Dialogschnittstellen. Das Projekt war durch drei Meilensteine gegliedert, die wie im folgenden beschrieben festgelegt wurden.

3.1.1 Meilenstein I - Herbst 1994

Die morphologische Analyse von Eingabestrings erfolgt mithilfe einer klassifikatorischen Morphologie, basierend auf Morphix. Als Ausgabe der Komponente wird eine getypte Struktur geliefert, die relativ zu einer Schnittstellendefinition in TDL interpretiert wird. Die Morphologie wird damit auf ca. 5.000 Lemmata und die entsprechende Zahl von Vollformen erweitert. Wenn ein Wort nicht erkannt wird, kann ein Klärungsdialog aufgerufen werden und das Wort in die Datenbasis aufgenommen werden.

Die syntaktische Analysekomponente kann über das bestehende Fragment hinaus auch komplexe Sätze der folgenden Art verarbeiten:

(1) Ich nehme an, daß du morgen keine Zeit hast.
(2) Den Termin, den du vorgeschlagen hast, müssen wir verschieben.
(3) Wir können uns treffen, bevor die Sitzung stattfindet.

Auf der semantischen und Diskurs-Ebene können Präsuppositionen in einer Semantiksprache repräsentiert werden. Auf der Basis dieser Repräsentation können die Anforderungen an das erforderliche Resolutionsverfahren und an das Diskursgedächtnis spezifiziert werden.

Die grammatische Verarbeitung erfolgt auf der Basis einer Integration der EBL-Methode mit dem Chart-Parser. Wenn ein Satz eine ähnliche Struktur wie ein bereits vorher von EBL gelernter Satz hat, kann er mithilfe von EBL analysiert werden. Wenn weitere Lösungen gefordert werden, die EBL nicht liefern kann, so analysiert der normale Parser unter Einbeziehung der von EBL gelieferten ersten Lösung. Es wird gezeigt, wie durch diese Strategie eine deutliche Performanzsteigerung der Analyse erreicht wird.

Das System verfügt über erste Mechanismen zur robusten Verarbeitung. Falls während der Verarbeitung aufgrund einer unvollständigen oder fehlerhaften Eingabe die weitere Verarbeitung unterbrochen wird, werden entsprechende Fehlerprotokolle aktiviert. Information über mögliche Fehlerquellen stammen von signifikanten Stellen in Morphologie, Lexikon und Parser. Diese Informationen werden von einer Monitoringkomponente in Abhängigkeit vom Gesamtzustand des Systems ausgewertet. Im Falle von nicht erkannten Wortformen wird zum Beispiel entweder ein interaktiver Klärungsdialog gestartet oder Defaultwissen benutzt. Kann

keine vollständige Parsinganalyse durchgeführt werden, wird mitgeteilt, welche Teilergebnisse berechnet und welche nicht kombiniert werden konnten.

Für die grammatikalische Generierung wird ein Prototyp einer top-down getriebene Earley-Deduktionsstrategie vorgeführt, der jedoch noch nicht in das Gesamtsystem integriert ist (siehe dazu Meilenstein II). Ausgehend von einer semantischen Merkmalsbeschreibung werden alle grammatikalisch zulässigen Wortketten erzeugt. Mittels Beispielgrammatiken wird die Funktionsweise der neuartigen Strategie gezeigt.

3.1.2 Meilenstein II - Herbst 1995

Die Scanning-Komponente ist in der Lage größere Texteinheiten, wie Absätze, Tabellen, Überschriften etc., zu erkennen. Dadurch kann die Makrostruktur eines Dokuments in die Verarbeitung miteinbezogen werden.

Die grammatische Abdeckung umfaßt den gesamten Bereich der nicht-finiten Konstruktionen, wie er in den folgenden repräsentativen Beispielen illustriert ist:

(1) Ich werde morgen keine Zeit haben.
(2) Ich hätte an dem Treffen teilnehmen können.
(3) Ich werde versuchen, rechtzeitig da zu sein.
(4) An dem Tag scheint ein Treffen stattzufinden.
(5) Das Treffen muß leider abgesagt werden.
(6) Dieser Termin ist voraussichtlich nicht einzuhalten.

Die Semantik- und Diskurs-Komponente verfügt über Methoden zur Auflösung von Referenz. Im Diskurs nicht explizit eingeführte Anaphern (Pluralanaphern, Ereignisanaphern, Nominalellipsen) können konstruiert werden. Die Menge der möglichen Antezendenten kann satzübergreifend erkannt werden und Präsuppositionen ausgewertet werden. Die unterschiedlichen semantischen Funktionen von Modalausdrücken werden erkannt und können in die verschiedenen pragmatischen Funktionen überführt werden.

Für die Integration des Kernsystems mit den in COSMA entwickelten Dialogteilen wird eine Schnittstelle definiert. Diese Schnittstelle legt die Arbeitsteilung zwischen linguistischer und nicht-linguistischer Verarbeitung fest und dient zum Austausch von entsprechenden Informationen.

Eine aufgaben- und daten-gesteuerte Architektur wird konzipiert und als erster Prototyp vorgeführt. Zur Realisierung einer integrierten Datensteuerung und zielorientierten Kontrolle wird das Konzept der getypten Datenspezifikation und der generischen Kontrollflußspezifikation vorgestellt. Anhand einer prototypischen Implementation eines Subsystems wird die Arbeitsweise dieser Architektur demonstriert. Für die relevanten Komponenten werden entsprechende getypte Schnittstellen spezifiziert. Die Auswahl der einzelnen Komponenten wird dann gleichmaßen durch den Zustand der aktuellen Eingabe als auch durch den spezifizierten generischen Kontrollfluß festgelegt.

Eine enge Verzahnung von Parser und EBL auf phrasaler Ebene wird vorgeführt. Für eine gegebene Wortkette werden mittels EBL alle möglichen anwendbaren phrasalen Muster bestimmt und als zusätzliche Eingabe dem Parser übergeben. Dieser benutzt diese Strukturen zur quasi-deterministischen Kontrolle. Dadurch wird die Effizienz gesteigert und gleichzeitig die Flexibilität erhalten. Es wird gezeigt, wie durch Einsatz statistischer Information (z.B. Zugriffshäufigkeit, Zugriffszeitpunkt) eine dynamische Re-Organisation und Kontrolle der EBL-Strukturen geleistet wird.

Die integrierte EBL/Parsing Methode wird zur Verarbeitung von Subsprachen eingesetzt. Auf der Basis von Textkorpora werden entsprechende Subsprachen-spezifische phrasale Muster in der Trainingsphase von EBL erworben. In der Anwendungsphase wird der Parser nur Zugriff auf diese Strukturen haben. Anhand der Auswertung von Testläufen wird das verbesserte Laufzeitverhalten des Parsers demonstriert.

Der Prototyp des Earley-deduktionsbasierten Generators wird als Teil des Gesamtsystems vorgeführt. Es wird gezeigt, wie durch Berücksichtigung von TDL-Typinformation und der dynamischen Typexpansion eine effiziente Steuerung realisiert ist.

Eine Strategie für das "Shallow Parsing" wird vorgeführt. Es wird gezeigt wie spezifische grammatikalische Strukturen (z.B. spezielle NP- und PP-Konstruktionen) sehr schnell verarbeitet werden können, wobei als Ergebnis getypte Merkmalsstrukturen geliefert werden. Die vorgeführte Methode ist noch nicht in das Gesamtsystem integriert (siehe dazu Meilenstein III).

Für das SADAW-Lexikon wird eine relationale Datenbank entwickelt, die die weitere Wartung und Pflege dieses Lexikons erleichtert. Für diese Datenbank wird eine Benutzer- und Entwicklerschnittstelle vorgeführt. Es wird exemplarisch gezeigt, wie man einen oder mehrere Stämme oder Lemmata nach bestimmten Kriterien sucht und diese dann sequentiell bearbeitet. Die Bearbeitung erfolgt mittels einer grafischen Benutzeroberfläche und ist weitestgehend mausgesteuert. Darüberhinaus ist es möglich Datensätze selektiv zu exportieren, d.h. ausgewählte Felder können nach vorgegebenen Kriterien exportiert werden. Die Datenbank ist zunächst noch nicht in das Gesamtsystem integriert (siehe dazu Meilenstein III).

3.1.3 Meilenstein III - Herbst 1996

Das in der Datenbank gespeicherte SADAW-Lexikon ist als Laufzeitversion in das Kernsystem integriert. Durch eine Schnittstelle zur Morphologie können über 120.000 Lemmata morphologisch verarbeitet werden. Das Lexikon ist darüberhinaus mit der grammatischen Verarbeitung über eine Schnittstelle verbunden, die die in SADAW enthaltene poly-theoretische, abstrakte Information in theorie-spezifische Strukturen überführt.

Bei der grammatischen Verarbeitung werden dadurch drei verschiedene Arten von lexikalischem Wissen verwendet. In einem Kernbereich von Funktionswörtern etc. werden unmittelbar in TDL spezifizierte Strukturen verwendet. Ein Grundwortschatz der häufigsten Wörter kann aus SADAW vorkompiliert und mit Information angereichert werden. Für den Restbereich kann SADAW als on-line Backup-Medium verwendet werden.

Für die demonstrierbare Funktionalität der Kernmaschine bedeutet das, daß in einem engeren lexikalischen Kernbereich ein tieferes grammatisches Verstehen gesichert ist, daß aber auch eine gewisse Robustheit in den lexikalischen Randbereichen gewährleistet ist.

Die Grammatik erfaßt Wortstellungsvariationen im Bereich der nominalen, präpositionalen und Sententialen Komplemente, sowie der Adverbiale, wobei Restriktionen, die sich aus Aspekten wie Pronominalisierung, Definitheit, grammatische Funktionen etc. ergeben, berücksichtigtwerden.

In Zusammenarbeit mit den in COSMA entwickelten Dialogkomponenten, kann das System Klärungsdialoge wie den folgenden verarbeiten:

A: Ich möchte mich mit Dr. Hinkelmann um 3 Uhr treffen.
B: Meinen Sie Dr. Knut Hinkelmann oder Dr. Elizabeth Hinkelman?
A: Es war Dr. Elizabeth Hinkelmann, mit der ich sprechen wollte.

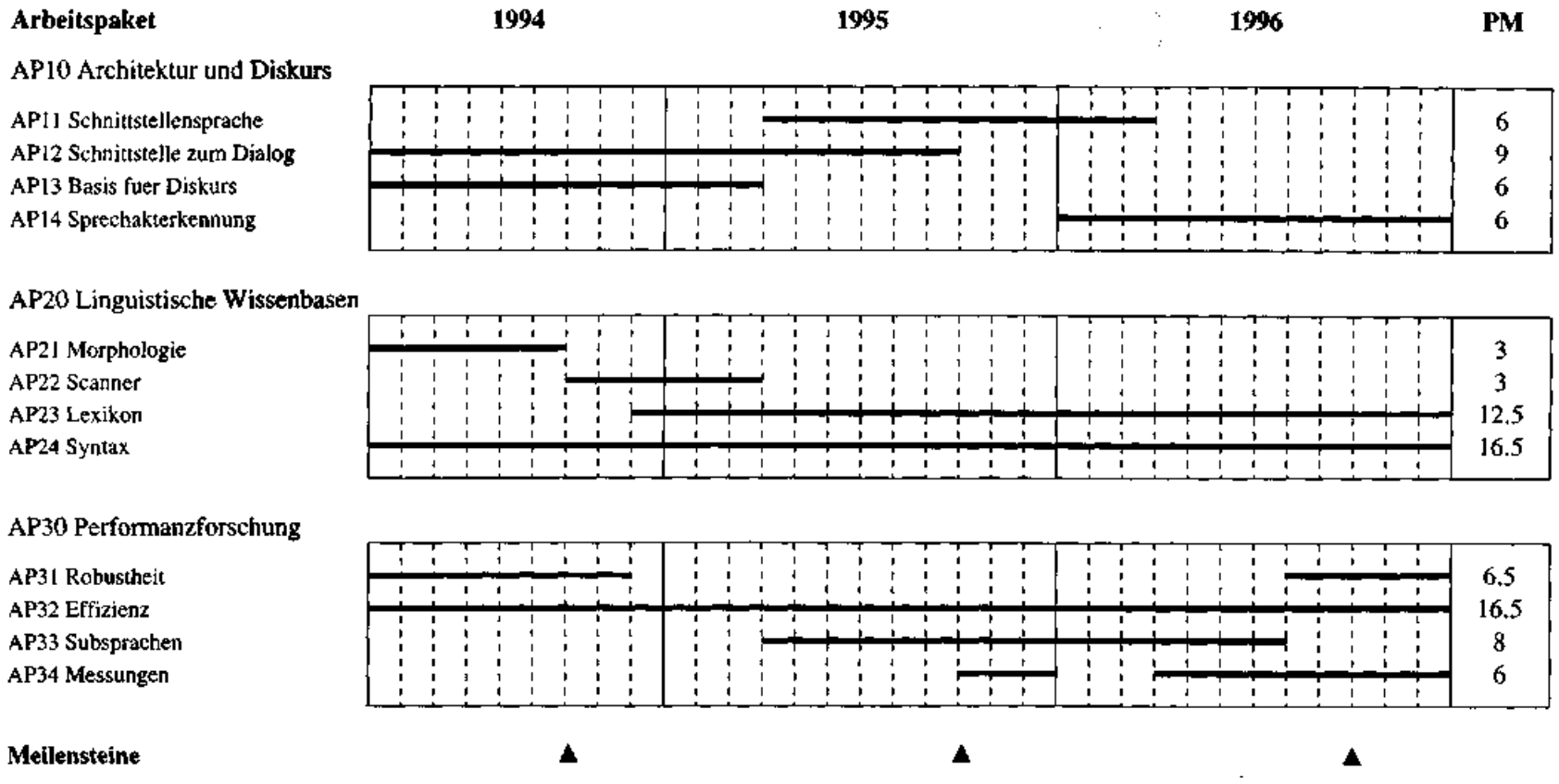
Dabei werden über die Schnittstelle zum Dialog linguistische und nicht-linguistische Informationen ausgetauscht, die zu einem Update des Kontextes und im Resolver führen.

Die Integration von Shallow Parsing, EBL und Parsing im Gesamtsystem wird vorgeführt. Ein gegebener String wird zuerst mittels Shallow Parsing analysiert und die gefundenen Teilergebnisse dem Parser übermittelt. Für nicht erkannte Reststrings wird die integrierte EBL/Parsing Komponente aktiviert und eine Gesamtanalyse der Eingabe durchgeführt. Falls keine vollständige Analyse möglich ist, werden die durch Shallow Parsing bereits verarbeiteten Teilstrings durch die EBL/Parsingkomponente erneut berechnet. Kann auch danach kein Gesamtergebnis berechnet werden, wird gezeigt, wie durch Verwendung von speziellen Klassifikationsschemata in einem nachfolgenden Schritt heuristisch orientierte Gesamtstrukturen gebildet werden.

Eine Methode zum automatischem Erwerb unbekannter Wörter basierend auf EBL wird vorgeführt. Enthält ein Eingabestring ein unbekanntes Wort, so wird mittels der EBL Datenstrukturen Information extrahiert, die zum automatischen Klassifizieren dieser Wörter herangezogen werden. Es wird gezeigt, wie mit dieser Methode unbekannte Nomina automatisch erworben werden können. Die Methode wird anhand ausgewählter Korpora evaluiert.

Die Verwendung von EBL für die Generierung wird vorgeführt. Es wird gezeigt, wie semantische Ausdrücke zum Indizieren von erworbenen Mustern benutzt werden. Für eine gegebene semantische Struktur werden zuerst entsprechende Muster extrahiert, die dann quasi-deterministisch auf entsprechende Wortketten abgebildet werden. Auch hier wird der Einsatz statistischer Information zur automatischen Reorganisation der EBL Strukturen demonstriert.

Balkenplan


Arbeitspaket
1994
1995
1996
PM
AP10 Architektur und Diskurs
AP11 Schnittstellensprache
AP12 Schnittstelle zum Dialog
AP13 Basis fuer Diskurs
AP14 Sprechakterkennung
6
9
6
6
AP20 Linguistische Wissenbasen
AP21 Morphologie
AP22 Scanner
AP23 Lexikon
AP24 Syntax
3
3
12.5
16.5
AP30 Performanzforschung
AP31 Robustheit
AP32 Effizienz
AP33 Subsprachen
AP34 Messungen
6.5
16.5
8
6
Meilensteine

3.2 Ablauf des Vorhabens

Zu Beginn des Projekts wurden die Projektziele gemäß den Auflagen des BMFT teilweise revidiert und in verschiedenen Punkten konkretisiert. Die Ergebnisse wurden dem BMFT als Ergänzungen zur Projektbewilligung im Juni 1994 schriftlich vorgelegt. Notwendig wurden die Änderungen vor allem infolge der um drei Monate verkürzten Projektlaufzeit. Die Präzisierung der Ziele erfolgte auch im Hinblick auf die Kürzungen im Vergleich zu dem Antrag, der dem Wissenschaftlichen Beirat vorgelegt worden war. Die Empfehlungen des Wissenschaftlichen Beirats wurden dabei berücksichtigt. Die Änderungen erforderten eine Neuabstimmung mit dem Schwesterprojekt COSMA, die entsprechend durchgeführt wurde.

Die Meilensteine, wie sie oben angegeben wurden, wurden im wesentlichen eingehalten und erfüllt. Es ergaben sich jedoch, u.a. durch die Arbeitsteilung mit dem COSMA-Projekt gewisse Akzentverschiebungen. So verschob sich zum Beispiel der Schwerpunkt von der Arbeit an der Dialog-Komponente in Richtung der flachen Verarbeitung, wie sie dann im Message Extraction System zum Einsatz kam. Von dieser Verschiebung profitierte am Ende auch das COSMA-System, in das die Message Extraction Komponente ebenfalls integiert wurde.

Die Arbeit an der Message Extraction Anwendung stellte eine Neubestimmung der geplanten zweiten Anwendung dar. Diese Neubestimmung wurde ebenfalls mit dem Wissenschaftlichen Beirat abgesprochen und von diesem begrüßt. Wie erwähnt, war ganz ursprünglich geplant, das PARADICE-System als interaktive Dialoganwendung, wie z.B. eine Datenbankabfrage, einzusetzen. Für diese Anwendung waren zwar keine Resourcen explizit veranschlagt worden, sondern es wurde davon ausgegangen, dass sich das System problemlos in eine solche Anwendung einbinden Hesse. Im Gegenzug und als Ersatz für diese Anwendung wurde nun das Message Extraction System SMES entwickelt, das innerhalb kürzester Zeit auch industrielles Interesse fand und die Grundlage des Nachfolge-Projects PARADIME darstellt.

Die Mitarbeiter im Projekt waren Dipl. Inf. Bernd Kiefer, Dr. Klaus Netter (Projektleitung), Dr. Günter Neumann und vorübergehend Gregor Erbach. Zu Beginn des Jahres 1995 erkrankte bedauerlicherweise der PARADlCE-Mitarbeiter Bernd Kiefer schwer und fiel während des gesamten Jahres aus. Dadurch bedingt kam es mitunter zu Verzögerungen bei der termingerechten Abwicklung bzw. zu Reduktionen des Umfanges von einigen Arbeitspaketen, die mit dem WBR abgesprochen wurden. Durch verstärkte Nacharbeit und die teilweise Verlagerung von Resourcen konnte der Gesamterfolg des Projektes jedoch sichergestellt werden.

4 Wissenschaftlich-Technischer Stand

Es existieren eine Reihe von große NL-Kernsysteme für das Deutsche. Die meisten von ihnen sind vom wissenschaftlichen Standpunkt aus überholt, selbst wenn man sie noch kommerziell für eine begrenzte Zeit verwenden kann. Firmen, die große Summen in Systeme investiert haben, die auf einer älteren Technik basieren, befinden sich häufig in einer Position, wo sie sich auf die Verwertung des Systems konzentrieren müssen, anstatt mit der Arbeit an einer neueren Technologie zu beginnen. Ein Beispiel ist Siemens-Nixdorf mit dem METAL MT-System, das möglicherweise das System mit der größten Abdeckung des Deutschen ist. Die Firma plant, die Kernkomponenten von METAL in zahlreichen verschiedenen Anwendungen zu verwenden. Das System besitzt keine Dialogfähigkeiten.

Zwei weitere Systeme für das Deutsche sollen hier erörtert werden. Der *Linguistic Kernel Processor* (LKP) von Siemens ist ein NL Kern-System, das sich seit 1987 in der Entwicklung befindet. In seinen Grundzügen ist es eng werwandt mit der *Core Language Engine* (CLE) vom SRI Cambridge. Genau wie bei CLE sind gewisse Kompromisse gemacht worden, um mit existierenden Beschränkungen bei Hardware und Software Effizienz zu erreichen. Der angewandte Formalismus ist eine idiosynkratische Entwicklung der NL-Gruppe bei Siemens. Es handelt sich um einen eher knappen unifikationsbasierten Grammatikformalismus, ausgestattet mit einigen speziellen Mechanismen, um Fernabhängigkeiten zu behandeln. Diese Mechanismen wurden beeinflußt von der linguistischen *Government and Bindung-Theorie* (GB). Die Prinzipien von GB sind jedoch nicht als solche ausgedrückt, sondern eher in große Mengen von Phrasenstruktur-Regeln übersetzt. Es ist vorwiegend der Fokus auf die Beschreibung von Phrasenstrukturen, der Geschwindigkeit zu erlangen hilft, da der Ergebnisreichtum vom Parsing von Phrasenstrukturen ausgenutzt werden kann. Der LKP ist mit das beste an linguistischer Technologie, was verfügbar ist. Er ist das Ergebnis von angewandter Forschung, die ausgerichtet ist auf kurzfristige Verwertung.

Das System LEU/2 von IBM Deutschland wurde gemeinsam mit fünf Universitäts-Partnern im Projekt LILOG gebaut. Das System wurde entwickelt für Aufgaben von Textverstehen. Dem NL-Kern liegt der Formalismus STUF-I zugrunde, ein merkmalsbasierter Unifikationsformalismus mit Disjunktion und parametrisierten Templates. Im Gegensatz zu LKP war die Entwicklung von LEU/2 erheblich mehr von längerfristigen Forschungszielen geleitet. Da man jedoch die Entwicklung während der letzten drei Jahre nicht fortgeführt hat, liegt LEU/2 bereits hinter dem derzeitigen Forschungsstand zurück.

Das PARADICE-Vorgängerprojekt DISCO begann kurz vor Ende des LILOG-Projektes. Da der Projektleiter von DISCO bei LILOG als Projektgruppenleiter und später als Leiter des Partnerprojektes an der Universität des Saarlandes gearbeitet hat, ist die wissenschaftliche Erfahrung aus LILOG von Beginn DISCO eingeflossen. In vieler Hinsicht hat DISCO Ansätze weitergeführt, die bei LILOG verwendet oder entwickelt wurden, und diese auf die Aufgabe angewendet, Dialogsysteme für kooperative Agenten zu bauen.

Es gibt international zahlreiche Projekte, die die eine oder andere Eigenschaft mit PARADICE gemeinsam haben. Da aber keines von ihnen auf die Entwicklung eines Kernsystems für das Deutsche abzielt, werden wir sie hier nicht weiter erörtern.

Es existiert kein System von der Größe des unsrigen, das mit einem höheren getypten merkmalsbasierten Unifikationsformalismus arbeitet, der dem jetzigen Forschungsstand entspricht. Ebenso sind uns weltweit keine Projekte bekannt, die das Ziel haben, ein parametrisierbares flexibles Diskurssystem mit Beschreibungssprachen höherer Ebene zu bauen.

Viele große Dialogsysteme für das Deutsche sind gebaut worden. Zwei bekannte Beispie-

le sind WISBER [HBB+88] und XTRA [AHK+89]. WISBER führt Beratungsgespräche über Geldanlagen. XTRA ermöglich natürlichsprachlichen Zugang zu Expertensystemen. Beide haben erheblich zum Fortschritt auf dem Gebiet des Modellierens von Dialogen beigetragen und eingehend an der Schnittstelle zwischen linguistischer Verarbeitung und Wissensverarbeitung gearbeitet. Beide verwenden weder moderne getypte merkmalsbasierte Unifikationsformal's-men noch folgen sie unserem Kompetenz-Performanz-System Ansatz. Sie sind vorwiegend darauf zugeschnitten, Dialog innerhalb einer bestimmten Klasse von Aufgaben abzudecken, wohingegen PARADICE sich um mehr Unabhängigkeit von Aufgaben bemüht.

Tatsache ist, daß uns keine ausbaufähigen Dialogsysteme bekannt sind, die Domänen-Unabhängigkeit gezeigt haben. Die offensichtliche Kandidaten für ausbaufähige Systeme sind das CEC PLUS-Projekt und SUNDIAL. Bisher hat PLUS das Problem der Interaktion zwischen Dialog und Domänen-Planung nicht gelöst. SUNDIAL hingegen ist gebunden an ein Dialogmodell, das auf domänenspezifischen Primitiven basiert. Es ist wahrscheinlich, daß seine allgemeine Technik - ein Markov-Modell - sich über ein große Zahl von einfachen Fällen als robust erweisen wird, bei komplexeren jedoch versagt.

5 Verwendete Fachliteratur

Literatur

[AHK+89] J. Allgayer, K. Harbusch, A. Kobsa, C. Reddig, N. Reithinger, and D. Schmauks. XTRA: A Natural-Language Access System to Expert Systems. *International Joint Man-Machine Studies*, 31:161-195, 1989.

[Bac92] Rolf Backofen. Using Distributed Disjunctions for Intelligent Backtracking, 1992. Paper presented at the Workshop on "Coping with Linguistic Ambiguity in Typed Feature Formalisms", ECAI 92.

[Bac93a] Rolf Backofen. On the decidability of functional uncertainty. In *Proc. of the 31 th ACL*, pages 201-208, Columbus, Ohio, 1993. Association for Computational Linguistics.

[Bac93b] Rolf Backofen. Regulär path expressions in feature logic. In Claude Kirchner, editor, *Proc. of the RTA '93*, pages 121-135, Montreal, Canada, 1993.

[BEG90] Rolf Backofen, Lutz Euler, and Günther Goerz. Towards the Integration of Functions, Relations and Types in an AI Programming Language. In *Proceedings of GWAI 1990*, 1990.

[BH93] Stephan Busemann and Karin Harbusch. DFKI Workshop on Natural Language Systems: Reusability and Modularity. Proceedings. Document D-93-03, DFKI, Saarbrücken, Germany, 1993.

[BKN+93] J. Bedersdorfer, K. Konrad, I. Neis, O. Scherf, J. Steffen, and M. Wein. Eine Spezifikationssprache für Transformationen auf getypten Merkmalsstrukturen. KI-93 Workshop: Neuere Entwicklungen der deklarativen KI-Programmierung, Berlin, 1993.

[BS93] Rolf Backofen and Gert Smolka. A complete and recursive feature theory. In *Proc. of the 31 th ACL*, pages 193-200, Columbus, Ohio, 1993. Association for Computational Linguistics. Extended version available as DFKI Research Report RR-92-30.

[EU90] Gregor Erbach and Hans Uszkoreit. Grammar Engineering: Problems and Prospects. CLAUS Report l, Computerlinguistik an der Universität des Saarlandes, 1990.

[HBB+88] Helmut Horacek, Henning Bergmann, Russell Block, Michael Fliegner, Michael Gerlach, Massimo Poesio, and Michael Sprenger. From Meaning to Meaning: A Walk Through WISBER's Semantic-Pragmatic Processing. In Wolfgang Hoeppner, editor, *Künstliche Intelligenz. Proc. GWAI-88, 12. Jahrestagung*, pages 118-129, Berlin, New York, 1988. Springer. IFB Vol. 181.

[HS92] Elizabeth A. Hinkelman and Stephen P. Spackman. Abductive Speech Act Recognition, Corporate Agents and the COSMA System. In W. J. Black, G. Sabah, and T. J. Wachtel, editors, *Abduction, Beliefs and Context: Proceedings of the second ESPRIT PLUS workshop in Computational pragmatics*, 1992.

[Kas93] Walter Kasper. Integration of Syntax and Semantics in Feature Structures. In Busemann/Harbusch 1993 [BH93J.

[Käs] Walter Kasper. Dynamic Interpretation of NLL. Research Report, DFKI Saarbrücken, to appear.

[KDD+92] Judith Klein, Ludwig Dickmann, Kader Diagne, John Nerbonne, and Klaus Netter. DiTo — Ein Diagnostikwerkzeug für deutsche Syntax. In Günter Görz, editor, *KONVENS 92,* pages 380-384. Springer, 1992.

[KN92] Hans-Ulrich Krieger and John Nerbonne. Feature-based inheritance networks for computational lexicons. In Ted Briscoe, Ann Copestake, and Valeria de Paiva, editors, *Default Inheritance Within Unification-Based Approaches to the Lexicon.* Cambridge University Press, Cambridge, 1992. A Version of this paper is also available as DFKI Research Report RR-91-31 .Also published in Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon, Technical Report No. 238, University of Cambridge, Computer La boratory, October 1991.

[KNP93] Hans-Ulrich Krieger, John Nerbonne, and Hannes Pirker. Feature-based allomorphy. In *Proc. of the 31 th ACL,* pages 140-147, Columbus, Ohio, 1993. Association for Computational Linguistics.

[KS93] Hans-Ulrich Krieger and Ulrich Schäfer. TDL — A Type Description Language for HPSG. Part 1: Overview. Technical Report, Deutsches Forschungsinstitut für Künstliche Intelligenz, Saarbrücken, Germany, 1993.

[Ner91a] John Nerbonne. Constraint-based semantics. In P. Dekker and J. van der Does, editors, *Proceedings of the 8th Amsterdam Colloquium,* 1991.

[Ner91b] John Nerbonne. Feature-based disambiguation. In Rod Johnson, Mike Rosner, and C.J.Rupp, editors, *Constraint Propagation, Linguistic Description and Computation,* 1991.

[Ner92] John Nerbonne. A feature-based syntax-semantics interface. In A. Manaster-Rama and W. Zadrozny, editors, *Proceedings of the 2nd International Conference on the Mathematics of Language,* 1992.

[Net92] K. Netter. On non-head non-movement. An HPSG treatment of finite verb position in German. In G. Görz, editor, *Proceedings of KONVENS 92,* Berlin and Heidelberg and New York, 1992. Springer.

[Net93a] Klaus Netter. Architecture and Coverage of the DISCO Grammar. In Busemann/Harbusch 1993 [BH93].

[Net93b] Klaus Netter. Towards a Theory of Functional heads. German nominal phrases. Lecture Note Series. CSLI, Stanford, 1993.

[Neu93] Günter Neumann. The DISCO Development Shell and its Application in the COSMA System. In Busemann/Harbusch 1993 [BH93].

[NLDO92] John Nerbonne, Joachim Laubsch, Abdel Kader Diagne, and Stephan Oepen. Natural Language Semantics and Compiler Technology. Research Report, Deutsches Forschungsinstitut für Künstliche Intelligenz, Saarbrücken, Germany, 1992.

[NNP93] J. Nerbonne, K. Netter, and C. Pollard, editors. *German Grammar in HPSG.* Chicago University Press, Chicago, 1993.

[NOD+93] John Nerbonne, Stephan Oepen, Abdel Kader Diagne, Karsten Konrad, and Ingo Neis. *NLL* — Tools for meaning representation. In Busemann/Harbusch 1993 [BH93].

[SH93] Stephen P. Spackman and Elizabeth A. Hinkelman. Corporate Agents. Technical Report, Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Germany, 1993.

[Tro90] Harald Trost. The application of two-level morphology to non-concatenative German morphology. In *Proceedings of the 13th International Conference on Computational Linguistics,* 1990.

[Usz91] Hans Uszkoreit. Strategies for Adding Control Information to Declarative Grammars. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* pages 237-245, 1991.

6 Zusammenarbeit mit anderen Stellen

6.1 Kooperationen innerhalb des DFKI

PARADICE arbeitet in sehr enger Kooperation mit dem COSMA-Projekt. Das Diskurssystem wird COSMA als Basis zum Experimentieren mit den Dialogparametern zur Verfügung gestellt. PARADICE liefert linguistische Kompetenzmodelle mit zunehmenden Ebenen der Robustheit, Effizienz und Flexibilität. COSMA wendet diese auf konkrete Beispiele an, um somit eine strenge Evidenz für ihre Nützlichkeit herzustellen. COSMA versieht PARADICE mit realistischen Kriterien für Funktionalität, Abdeckung und Performanz. Das Konzept der Subsprachen, das PARADICE implementiert, wird auch von COSMA angewendet und getestet.

Im Bereich Messung und Entwicklung von Testkorpora (AP 34) arbeitet PARADICE mit dem LRE-Projekt "Test suites for NLP applications" zusammen, das unter anderem auf dem am DFKI entwickelten DiTo-System basiert. Die Partner des DFKI in diesem Projekt sind die Universität von Essex, ISSCO (Genf) und AEROSPATIALE (Frankreich). Die Ergebnisse dieses Projektes sind direkt für das Vorhaben verwendbar.

Die Kooperation zwischen DISCO und WIP hinsichtlich der Kompilation von HPSG-Grammatiken in TAG-Bäume für die effiziente Generierung wird in den jeweiligen Nachfolgeprojekten fortgesetzt.

6.2 Kooperationen mit den Gesellschaftern

Der Kontakt mit Siemens durch Dr. U. Block, dessen Team eine große Grammatik des Deutschen entwickelt hat, wird fortgesetzt. Im Rahmen der Zusammenarbeit mit Siemens stehen auch die Arbeiten in AP 23 zur Wiederverwendung und Nutzbarmachung des SADAW-Lexikons. (In diese Zusammenarbeit ist auch das IMS an der Universität Stuttgart einbezogen.) Obwohl PARADICE eine andere Orientierung als das Siemens-Projekt hat und obwohl die Projekte sich in ihrem Formalismus unterscheiden, gibt es gemeinsame Interessen in der Erweiterung von Grammatik und Lexikon.

Enge Zusammenarbeit existiert mit der Gruppe bei IBM Deutschland im Bereich des Formalismus. Durch die gemeinsamen Arbeit an der Grammatik in VERBMOBIL, besteht eine Kollaboration im Bereich des Grammatikentwicklung und der linguistischen Wissensbasen.

Seit längerem bestehen Kontakte zwischen der Disco-Gruppe und Daimler-Benz. Dies hat schließlich zu einem Tandem-Projekt zwischen Daimler und den DFKI-Gruppen von Prof. Wahlster und Prof. Uszkoreit geführt. Im Projekt EFFENDI, das von Daimler finanziert wird, werden Generierungstechniken implementiert, die Daimlers Arbeit über Analyse komplementieren. . . . .

6.3 Kooperation mit VERBMOBIL

Das PARADICE-Projekt hat eine einheitliche Sichtweise, die theoretische Kohärenz, uniformen Formalismus und eine flexible Systemarchitektur umfaßt. Aufbauend auf den Leistungen von DISCO in Formalismus und linguistischen Kompetenzmodellen versucht PARADICE, signifikante Fortschritte in linguistischen Performanzmodellen in Beziehung zu Kompetenz, und im parametrisierbaren Dialog zu erreichen. Diese Ergebnisse werden in einem allgemeinen und wiederverwendbaren Softwaresystem integriert. Im Gegensatz dazu verfolgt das VERBMOBIL-Projekt die Integration von geschriebener und gesprochener Sprache für spontane Dialoge

innerhalb eines bestimmten Anwendungskontextes, nämlich der automatischen Übersetzung. In diesen gemeinsamen Bereichen der beiden Projekte ergeben sich starke Synergieeffekte.

Synergetische Beziehungen zwischen VERBMOBIL und PARADICE gibt es in allen Bereichen von Grammatikentwicklung, effizienter Verarbeitung von constraintbasierten Formalismen, Semantik und Dialogverarbeitung. Enge Zusammenarbeit und Austausch von Ergebnissen zwischen VERBMOBIL und PARADICE ist durch die Tatsache garantiert, daß VERBMOBIL-Forschung in den relevanten Bereichen vom DFKI selber (zum Teil durch Forscher des ehemaligen ASL-Projektes) und am Lehrstuhl für Computerlinguistik an der Universität des Saarlandes (Prof. Pinkai) durchgeführt wird.

Im Bereich der Grammatikentwicklung verfolgt VERBMOBIL zwei Strategien. Für die Verarbeitung spontaner gesprochener und geschriebener Sprache baut man auf den Erfahrungen und Ergebnissen des Siemens LKP auf, welches sofort schon eine technologisch ausgereifte Umgebung und ein umfassendes Grammatikfragment zur Verfügung stellt. Parallel dazu wird eine grundsätzlich neue Implementierung einer HPSG-artigen Grammatik für das Deutsche entwickelt, für die IBM die Verantwortung übernimmt.

Die Grammatikentwicklung in PARADICE profitiert auf der einen Seite von den Ergebnissen in VERBMOBIL im Bereich gesprochener Sprache. PARADICE stellt jedoch auch ein state-of-the-art HPSG-Grammatik Fragment für das Deutsche zur Verfügung mit weitergehenden Erweiterungen in verschiedenen relevanten Aspekten linguistischer Wissensbasen. Das Basisfragment und die Erweiterungen werden VERBMOBIL zur Verfügung gestellt. Das existierende Fragment der DISCO-Grammatik ist auf die Anwendungsdomäne der Terminplanung eingerichtet und deshalb eng verbunden mit der Domäne der Konferenzabsprachen, die in VERBMOBIL ins Auge gefaßt wird.

PARADICE basiert auf dem fortgeschrittenen Formalismus, der im DISCO-Projekt entwickelt und extensiv getestet wurde und unternimmmt deswegen selbst keine neuen Anstrengungen in der Entwicklung neuer constraintbasierter Formalismen. Das BMFT hat bereits in dem Zuwendungsbescheid für VERBMOBIL das DFKI gebeten, den TDL-Formalismus in VERBMOBIL zur Verfügung zu stellen. PARADICE wird zusätzlich Techniken aus dem Bereich der effizienten Verarbeitung für diese Formalismen liefern, die von direktem Interesse für VERBMOBIL sind.

Arbeiten in der Semantik in VERBMOBIL sind essentiell für PARADICE, da dort keine Mittel für dieses Gebiet zur Verfügung veranschlagt sind. Im besonderen werden die Arbeiten im Bereich der Dialogsemantik und Ellipsen, die in den DFKI Verbmobil-Teilprojekten durchgeführt werden, übernommen und an dasPARADICE-Systemangepaßt. Da in der Anfangsphase von VERBMOBIL keine stabile Testumgebung existiert, dient das DISCO- (oder PARADICE-) System weiter als Testumgebung für die Semantikarbeiten der ASL-Nachfolgeprojekte in VERBMOBIL am DFKI.

6.4 Kooperation mit anderen Partnern

- Der Austauch von Ideen mit dem Lehrstuhl für Computerlinguistik and der Universität des Saarlandes hat sich als fruchtbar erwiesen und wird ohne Zweifel weiterbestehen. Das BiLD-Projekt (Bidirectional Linguistic Deduction) und PARADICE sind beide an der Modellierung von Wissen interessiert, das sowohl für die Erkennung als auch für die Generierung benutzt werden kann.

- Die PARADICE-System wird am CSLI (Stanford University) für die Gruppe von Prof. Ivan

Sag und Dr. Daniel Flickinger (Verbmobil) installiert. Von den Forschern am CSLI wurde dieses System äußerst positiv aufgenommen und wird zur Entwicklung einer englischen HPSG-Grammatik eingesetzt. Über das CSLI entstand auch ein Kontakt zur Simon-Frazier Universtität in Vancouver, wo das PARADlCE-System ebenfalls installiert wurde und zur Entwicklung einer japanischen HPSG und zu Unterrichtszwecken verwendet wird.

- Das ERGO (English Resource Grammar On-Line) Konsortiums entscheidet sich einhellig, für die verteilte Grammtikentwicklung in ERGO das (u.a.) in PARADICE entwickelte PAGE-System (Platform for Advanced Grammar Engineering) einzusetzen. Dadurch entstehen wissenschaftliche Kontakte zu allen in dieser Gruppe organisierten Institutionen, wie z.B. Simon-Fräser University, Ohio University, Carnegie Mellon University, u.a.

- Das TDL-System wurde Prof. James Pustejowsky von der Brandeis University, Mass., zur Verfügung gestellt. Es wird dort auf die Verwendbarkeit in einem Projekt zur lexikalischen Semantik (in Kooperation mit der Firma Apple) evaluiert. Diese Zusammenarbeit wurde durch einen Besuch von Dr Michael Johnston von der Brandeis University noch weiter ausgebaut.

- Prof. Vijay-Shanker von der University of Delaware verbringt 1994 sein Sabbatical an der DFKI GmbH. Zusammen mit ihm wurde die theoretische Basis und ein Algorithmus für die Compilierung von HPSG- in TAG-Grammatiken entwickelt. Diese Arbeit wurde unterstützt durch einen kurzen Gastaufenthalt von Prof. Robert Kasper (Ohio State University). Im Rahmen eines Workshops zum Thema "HPSG and TAGs" am DFKI, Saarbrücken (12./13. 06. 94), kam es zu weiteren intensiven Kontakten mit Prof. Aravind Joshi und Prof. Anthony Kroch (Univ. of Pennsylvania), sowie Dr. Anne Abeille und Dr. Owen Rambow (Universite Paris VII). Klaus Netter verbringt 1995 einige Wochen auf Einladung von Prof. Joshi als Gastwissenschaftler am Institute for Research in Cognitive Science (IRCS) der University of Pennsylvania, Philadelphia.

- Zusammen mit der Berteismann AG, dem italienischen Software Konzern Datamat, und der Belgischen Niederlassung von TRADOS wird ein Antrag auf ein Telematics Language Engineering (TAP-LE) Projekt im Bereich Multilingual Indexierung und Vernetzung von WWW-Dokumenten eingereicht (MULINEX). Das Projekt beginnt im September 1996 und baut u.a. auf die SMES-Technologie aus PARADICE auf. Unabhängig von diesem Projekt bestehen zur Berteismann AG darüberhinaus Kontakte im Bereich von Call-Center-Technologien.

- Im Rahmen des EU-Projekts TWENTY-ONE (TAP, IE, Beginn Januar 1996), das die Erschließung, die Dissemination und das Retrieval von Dokumenten im Umweltschtuzbereich zum Thema hat, bestehen Kontakte u.a. zu TNO Delft, der Niederländischen Organisation für Angewandte Forschung, zu Prof. Franciska de Jong an der Universität Twente, sowie dem Europäischen Forschungszentrum von XEROX in Grenoble.

- Zusammen mit TNO Delft, der Universität Twente und einer Reihe von größeren Fernsehstationen (TROS, NL; BRTN, Belgien; Südwestfunk Baden-Baden) wird das EU-Projekt POP-EYE im Language-Engineering-Sektor eingereicht, das als Ziel die multi-

linguale Indexierung und das Retrieval von Video-Material auf der Basis von Untertiteln hat. Dieses Projekt beginnt im Januar 1997 und baut auf Ergebnisse von PARADICE auf.

- Aus der Zusammenarbeit im TSNLP-Projekt und den Arbeiten in PARADICE geht später ein weiteres EU-gefördertes TAP-LE Projekt hervor (DiET - Diagnostic and Evaluation Tools for Natural Language Applications). Dieses Projekt beginnt Anfang 1997 und wird gemeinsam mit den TSNLP-Partnern Aerospatiale und ISSCO, sowie der IBM Deutschland, SRI Cambridge und dem Irischen Software Localization Resources Centre durchgeführt.

7 Erzielte Ergebnisse

Das Projekt gliederte sich in drei Bereiche, Architektur und Diskurs, Linguistische Wissensbasen, und Verarbeitung. Während der erste Bereich sich im Wesentlichen auf die Integration, die tieferen Verarbeitungsebenen und die Schnittstellen zu den Anwendungen konzentrierte, wurden im zweiten Bereich die linguistischen Resourcen entwickelt, die als Grundlage für die Verarbeitung dienten.

7.1 Architektur und Diskurs

Der Aufgabenschwerpunkt im Bereich Architektur und Diskurs lag auf der einen Seite auf der Entwicklung einer flexiblen Systemarchitektur und Integration, auf der anderen Seite auf der Entwicklung von diskursspezifischen Teilen, die u.a. für die Integration in den COSMA-Prototypen relevant waren.

AP 11: Schnittstellensprache

Im Bereich der Schnittstellensprache wurde ein Ansatz fuer eine neue Datenstruktur für datengetriebene Textverarbeitung entwickelt. Die Grundidee baut dabei auf das Konzept der Chart auf, wie sie aus dem Bereich des Parsens einzelner Sätze bekannt ist, und erweitert sie in Richtung von Texten. In einer sogenannten Textchart werden die verschiedenen Ebenen der Textanalyse über ihre Positionen im Text zu einander in Beziehung gesetzt. Dies erlaubt gleichzeitig einen simultanen Zugriff zu Informationen auf verschiedenen Textebenen, wie auch ein komponentenspezifisches Filtern der Information.

Der Schwerpunkt in diesem Paket lag gegen Ende des Projektes auch auf der Integration der Kernmaschine und des Message Extraction Systems. Beide Systeme verfügten am Ende über eine gemeinsame Menge von Kernkomponenten, speziell das Typ-System TDL, den Unifikator, die Morphologiekomponente, und das Lexikon (incl. der morphologischen Typdefinitionen der Syntax-Morphologie-Schnittstelle aus der HPSG Grammatik). Es wurde eine erste Version der Schnittstelle zwischen der flachen und der tiefen Analyse implementiert, mit der resultierende Ausdrücke der flachen Analyse dem Parser der tiefen Analyse als quasi komplexe Lexikoneinträge übergeben werden (analog der Schnittstelle zwischen EBL und Parser). Damit kann die Chart des Parser bereits mit flachen Ausdrücken initialisiert werden. Weiterhin wurde das Message Extraction System in das COSMA System integriert.

AP 12: Schnittstelle zum Dialog:

Das Ziel des Arbeitspakets war es, den Ansatz der Verwendung eines uniformen Formalismus auf verschiedenen Ebenen zu erweitern und zu verbessern, um dadurch eine adäquate UnterstützungfürdieDialogverarbeitungzuerzielen,dieüberdietraditionellenParsing-undGenerierungsprozessehina

Im Bereich der Schnittstelle zwischen Dialogmodellen und anderen Formalismen wurden zunächst beispielhafte Kodierungen von typischen Repräsentationsproblemen konstruiert, die u.a. Agenten-Modelle, zeitliche Sequenzen, interpretierte, propositionale Inhalte und Fehler-Protokolle für vorausgehende linguistische Module (akustisches Fehlverstehen, Mißverstehen) umfassen. Diese Arbeiten zeigten die Notwendigkeit, in den Formalismus einen Begriff von

theoretischer *Präferenz* zu integrieren, der in engem Zusammenhang zu Default-Theorien steht.

Dieser Begriff ist distinkt sowohl von Verarbeitungs-Pn'oniäten, wie sie im PARADICE-System eingesetzt werden, um die Parsingstrategie und (teilweise) auch Unifikation zu steuern, als auch von empirischen *Wahrscheinlichkeiten,* die zum Beispiel auf der Basis von Trainingsdaten berechnet werden können. Solche Präferenzen würden auf der deskriptiven Ebene die Erfassung einer Reihe von Dialog- und anderen generellen Phänomenen unterstützen. Es wurde gezeigt, daß Ansätze, die die Verarbeitungsprioritäten zu diesem Zwecke ausnutzen, zu einer Verschlechterung der Systemperformanz führen können.

Eines der wichtigsten Ergebnisse war dabei auch, daß die Dialog-Grounding-Funktionalität weitgehend direkt im existierenden TDL-Formalismus beschrieben werden kann. Da dieser Formalismus Turing-Mächtigkeit besitzt, umfassen die Ausnahmen zu dieser direkten Beschreibung innovative Kodierungstechniken, Erweiterungen der operationalen Semantik, als auch Erweiterungen des Formalismus für bequemere und einfachere Beschreibungen. Diese Erweiterungen spiegeln sich auch in den folgenden Punkten, die die Gesamtresultate zusammenfassen:

(i) eine nahezu vollständige Kodierung der Dialogfunktionalität in getypten Merkmalsstrukturen;

(ii) die Entwicklung einer listen-basierten Kodierung von FSM für die Modellierung von Sequenzen von Grounding Acts, die einen wohlgeformten Kern-Sprechakt darstellen;

(iii) die Entwicklung eines Konzepts von semantischen Präferenzen zur Behandlung von Aspekten der Reversibilität und Intentionserkennung;

(iv) eine verbesserte und robustere Behandlung von Asynchronie und des Managements von Dialogzuständen.

Obzwar noch eine ganze Reihe von offenen Forschungsfragen verbleiben, hat sich gezeigt, daß getypte Merkmalsformalismen ausreichend erweiterbar sind, um typische Phänomene im Dialogbereich zu erfassen.

AP 13: Basis für Diskurs

Im Bereich der Syntax-Semantik-Schnittstelle wurde die Übersetzung der semantischen Merkmalsstrukturen nach NLL auf ein typgesteuertes Verfahren umgestellt, statt einen eigenen Klassifikator zu benutzen. Dies erhöht die Transparenz und erleichtert die für die Generierung erforderliche Übersetzung von logischen NLL-Strukturen in semantische Merkmalsstrukturen.

Bei der Verarbeitung wurden*backtracking-Mechanismen*für alle Resolutionsprozesse implementiert, um alternative Lösungen zugänglich zu machen. Diese Mechanismen sind in die PARADICE-Architekturintegriert.

Das zur Referenzauflösung verwendete Diskursgedächtnis wurde um eine Wissensbasis für Fakten mit einer eigenen PROLOG-artigen Inferenzmaschine erweitert, die für die Resolution einfacher und komplexer Kennzeichnungen eingesetzt wird. Auch wurden Schnittstellen vom NLL-System zur Sortenlogik von TDL implementiert, um für den sortalen Abgleich während der Resolution die Ontologie der Merkmalssemantik verfügbar zu machen. Daneben wurden zusätzliche Mechanismen zur Einbeziehung konzeptuellen Wissens für die Referenzauflösung integriert, insbesondere Teil-Ganze-Beziehungen.

Der Leistungsumfang der Resolutionskomponente wurde in Richtung auf die Behandlung

von Nicht-Identitäts-Anaphern erweitert. Dabei wurden zunächst ein Summationsverfahren für Pluralanaphern und Abstraktionsverfahren für N-bar-Anaphern realisiert. Das Diskursgedächtnis wurde auf die Dialogverarbeitung vorbereitet, indem Mechanismen zur Verfolgung von Sprecher-Hörer-Wechseln integriert wurden. Am Beispiel von Terminen wurde als zusätzliche Komponente das Konzept des *Diskurs-Topics* eingeführt, welches zusammen mit einer in Entwicklung befindlichen Komponente zur Terminverfolgung in Terminvereinbarungsdialogen ein abstraktes Referenzobjekt in Fällen liefert, in denen ein direkter Match mit Antezedenten nicht möglich ist (Beispiel: *Ich schlage den Montag vor.* - Das *geht bei mir nicht.).* Dazu wurden auch Konsistenztests für Zeitangaben untersucht und implementiert. Dabei wurde die Resolutionskomponente für die Resolution temporaler Ausdrücke und die Konstruktion kontextueller temporaler Referenzpunkte erweitert. Parallel dazu wurde begonnen, Lokalanaphorik einzubeziehen, wobei zunächst eine Beschränkung auf explizit eingeführte Lokalitäten besteht. Damit wurde eine dritte wichtige Phänomenklasse im Bereich der Nicht-Identitätsanaphern bearbeitet.

Darüberhinaus wurden Schritte unternommen, um die Resolution mit der Aufgabe der Sprechaktidentifizierung zu verknüpfen. Im Rahmen dieser Arbeiten wurde auch die Interaktion mit linguistischen Sprechaktindikatoren wie Modalausdrücken und performativen Verben untersucht, die in Erweiterungen des semantischen Sortensystems, d.h. der semantischen Ontologie, resultierte.

AP 14: Sprechakterkennung

Die Sprechakterkennung wurde in das Kernsystem integriert. Grundlage ist ein domänenunabhängiges Sprechaktklassifikationsschema. Die Sprechakte sind in einem Sortenverband im TDL-Formalismus definiert. Zur Unterstützung dieser Arbeiten wurden Möglichkeiten untersucht, die semantische Analyse von Sprechaktindikatoren, etwa performativen Phrasen und Diskurspartikel, zu erweitern. Feste Phrasen, wie *guten Tag* und performative Ausdrücke werden bereits im Lexikon als Sprechakte markiert. Die Erkennung wurde in die Resolutionskomponente integriert. Damit ist eine Interaktion mit der Referenzaufiösung möglich. Diese wurde zugleich um Zugänglichkeitsconstraints für Antezendenten in Frageakten ergänzt, wobei allerdings die Interaktion mit der Diskursstruktur noch weiter spezifiziert werden muß.

Der Sprechakterkenner wurde in Richtung eines möglicherweise auch domänenspezifischen Diskursmodells parametrisiert. Für die Spezialisierung des Erkenners in Bezug auf diese Modelle wurde eine Makro-Sprache entwickelt, die es erlaubt, sprechakt-bezogene semantische Constraints für die Erkennung auszudrücken, die dann Teil des Diskursmodells werden können.

7.2 Linguistische Wissensbasen

Im Bereich linguistische Wissensbasen lag das Ziel vor allem darin, umfangreiche wiederverwendbare Resourcen zu schaffen, bzw. bestehende Resourcen zu erschließen, auf den aktuellen Stand der Technik zu bringen und in das System zu integrieren. Für diejenigen Resourcen, die auch als unabhängige Module sollten eingesetzt werden können, waren deshalb flexible und systemübergreifende Schnittstellen, sowie benutzerfreundliche Wartungsmöglichkeiten wichtige Kriterien.

AP 21: Morphologie

Da abzusehen war, daß die in DISCO verwendete Morphologiekomponente X2MORF für die Verarbeitung von größeren Datenmengen nicht über eine ausreichende Robustheit und Effizienz verfügt, wurde diese Komponente durch das Morphologiesystem MORPHIX ersetzt. Der Nachteil, daß MORPHIX, im Gegensatz zu X2Morf, nicht mit den in der Grammatik verwendeten TDL-Datenstrukturen arbeitete, mußte bei der Integration in das Gesamtsystem durch eine Schnittstelle ausgeglichen werden. Das Problem dabei war vor allem, daß das MORPHIX-System und die Grammatik unterschiedliche Merkmale und auch unterschiedliche Merkmalskonfigurationen vorsah. Eine direkte Abbildung zwischen diesen Merkmalen wäre mit einem erheblichen Aufwand verbunden gewesen und hätte aus wartungstechnischen Gründen diese spezifischen Merkmalskonfigurationen praktisch für alle Zeiten festgelegt.

Als ein neuartiger Typ von Schnittstelle wurden deshalb im TDL-Formalismus morphosyntaktische Typen spezifiziert, die alle möglichen Ausgaben der Morphologie in Äquivalenzklassen aufteilen, unabhängig davon, welches die morphologische Klasse des zugrundeliegenden Lexems ist. Diese Äquivalenzklassen decken alle möglichen morphologischen Formen und paradigmatischen Merkmalskonfigurationen des Deutschen ab und definieren damit eine extensionale Schnittstelle. Jeder Analyse-Output von MORPHIX kann so durch eine einfache Vergleichsoperation auf einen Klassen- oder Typnamen abgebildet werden, der dann im TDL-System expandiert werden kann.

Aus diesem Ansatz ergeben sich mehrere Vorteile. Zum einen liegt die Definition der morpho-syntaktischen Merkmalskonfigurationen nunmehr vollständig im Bereich der Grammatik, da die Typdefinitionen jederzeit unabhängig verändert werden können, solange die Klassen- oder Typnamen beibehalten werden. Dadurch verfügt die Grammatik über eine getypte Schnittstelle, die die genauen Anforderungen an eine Morphologiekomponente spezifiert, so daß theoretisch auch andere Morphologiesysteme angeschlossen werden könnten. Zum anderen erhielt MORPHIX dadurch eine Schnittstelle, die die Verbindung zu anderen Verarbeitungssystemen wesentlich vereinfacht und auch die Möglichkeit bietet, MORPHIX als morpho-syntaktischen Tagger einzusetzen. Daraus ergibt sich auch, daß die Schnittstelle das Potential für eine theorie-unabhängige Standardisierung in sich trägt, die in dieser Form zuvor nicht möglich gewesen wäre.

Das morphologische Lexikon von MORPHIX wurde erheblich erweitert, nämlich um ca 80.000 neue Nomen- und 25.000 neue Adjektiveinträge und ca. 15.000 Verben Ausgangspunkt für die Erweiterung waren entsprechende Einträge aus dem SADAW-Lexikon. Die neuen MoRPHlX-Einträge wurden dabei automatisch mit Hilfe einer Übersetzungsroutine erzeugt, die im wesentlichen eine Transformation der morphosyntaktischen Information von der SADAW-Kodierung in die entsprechende MORPHIX-Kodierung durchführt. Insgesamt verfügt MORPHIX damit nun über mehr als 120.000 Stammeinträge.

Bei der Verarbeitung dieses wesentlich erweiterten Lexikons zeigte sich, daß die bis dahin verwendete Speicherung der lexikalischen Einträge mittels Hash-Tabellen zu einem erheblichen Effizienzverlust führte. Um diesen Verlust an Performanz wieder aufzufangen, wurde MORPHIX auf die Verwendung von TRIEs (Buchstabenbäumen) angepaßt. Die spezifische Implementation der TRIEs verhinderte, daß das neue, wesentlich größere Lexikon einen Verlust der Performanz zur Folge hatte, sodaß der Zugriff nahezu konstant bleibt. Neben den üblichen Operationen auf TRIEs (Einfügen, Zugreifen, Löschen) wurden auch Operationen implementiert, die es erlauben, Information existierender Einträge zu überschreiben und zu erweitern, wobei dies auch für beliebige, nicht MORPHIX-spezifische Information gilt. Des weiteren wurden Operationen definiert, die Zugriffe mittels Mustern möglicher Lemmata erlauben.

Über den ursprünglichen Arbeitsplan hinausgehend wurde die Morphologiekomponente MORPHIX um eine effiziente Behandlung von nominalen Komposita erweitert. Dieser Ansatz berücksichtigt unter anderem die möglichen Fugenelemente bei Nominalkomposita auf der Basis linguistisch fundierter Genereralisierungen. Durch diese Erweiterung konnte die Abdeckung der Morphologie noch einmal wesentlich gesteigert werden. Erste Performanztests mit einem größeren Korpus von freiem Text, einem Lexikon von ca. 80.000 Stämmen und einer Coverage von ca. 90 % zeigten, daß mit einer Verarbeitungszeit von unter 3 msec pro Wort (inkl. aller Lesarten) die Geschwindigkeit durchaus industriellem Standard entspricht.

AP 22: Text Scanning

Bei der Verarbeitung von natürlichsprachlichen Texten und Dokumenten spielt das Einlesen und die Normalisierung der Texte eine wichtige Rolle. Andererseits existiert bislang keine generalisierbare und anwendungsunabhängige Lösung für dieses Problem, auch wenn die Standardisierung durch SGML sich möglicherweise als ein Schritt in diese Richtung erweisen könnte. Aus diesem Grunde, wurde dem Text-Scanner bzw. der Text-Layout-Erkennung im Projekt nur eine stark untergeordnete Rolle zugewiesen.

Der Scanner wurde allerdings zumindest soweit erweitert, daß ganze Dateien (und nicht nur einzelne Sätze) verarbeitet, erkannt und markiert werden können. Dadurch kann z.B. eine Datei in eine Liste von Paragraphen aufgespalten werden, die wiederum aus Listen von Sätzen bestehen. Schlüsselwörter in tabellarischen Auflistungen können speziell behandelt werden und spezielle Schreibweisen, wie z.B. gesperrt gedruckte Worte oder Trennstriche, können erkannt werden.

AP 23: Lexikon

Ein Schwerpunkt im Projekt war es, die Resourcen des SADAW-Lexikons in eine Lexikon-Datenbank mit maschinell verarbeitbarer Informationsstruktur zu überführen. Als Resultat entstand eine Lexikon-Datenbank und Lexikographen-Workbench, SARDIC, die sowohl vom Inhalt, als auch von der Struktur und der Bedienungsoberfläche modernsten Standards genügt.

In einer ersten Phase wurde eine Konzeption von Datenstrukturen entwickelt, die sich einerseits für den Benutzer in transparenter, leicht verständlicher und linguistisch moderner Form darbieten, die aber in der internen Darstellung den Anforderungen eines intelligenten relationalen Datenbanksystems folgen. Des weiteren wurde die Oberflächenfunktionalität definiert, sowie die Strukturen für den Datenimport und -export festgelegt.

Diese Konzeption wurde in einer zweiten Phase zu einem Prototyp einer Lexikon-Workbench ausgebaut und verbessert. Im einzelnen wurden dabei folgende Arbeiten ausgeführt:

(i) Die Konzeption des Datenbankschemas für Nomina und Adjektive wurde implementiert und die Daten aus der alten SADAW-Datensammlung importiert. Zur weiteren Bearbeitung dieser Daten wurden mehrere Tools, sowie eine prototypische graphische Benutzerschnittstelle definiert. Es wurde ein Tool implementiert, das es erlaubt, Nominalkomposita halbautomatisch auf ihr Grundwort zurückzuführen. Mithilfe dieser Information kann die Konsistenz von Einträgen überprüft und verbessert werden, indem Abweichungen zwischen der Kodierung (einer Gruppe) von Komposita und der Kodierung des dazugehörigen Grundwort identifiziert werden.

Die Analyse und Korrektur der Nominalkomposita wurde mithilfe dieser Tools durchgeführt, und es wurden alle in der Datensammlung vorkommenden Komposita zerlegt und soweit vorhanden mit dem entsprechenden Simplex (oder Grundwort) verbunden. Bei fehlendem Simplex wurde dieses eingefügt. Die zahlreichen Fehler in der Kodierung, die durch diese Verbindung festgestellt werden konnten, wurden beseitigt. Die Datenbank wurde so erweitert, daß für dasselbe Grundwort (oder Lemma) auch verschiedene Morphologien (verschiedene Artikel, Deklinationen, Pluralbildungen) oder Lesarten eingetragen werden konnten. Bei der Zuordung von Komposita zu Simplizia (z.B. bei *Zentralbank, Gartenbank)* wurde diese Mehrdeutigkeit entsprechend berücksichtigt.

(ii) Es wurde damit begonnen, die SARDIC-Daten auch mit anderen umfangreicheren, maschinenlesbaren lexikalischen Ressourcen im Hinblick auf Abdeckung und Konsistenz zu vergleichen. Um eine Vergleichbarkeit zu ermöglichen, mußte unter anderem eine Normalisierung der Lemmata, bzw. eine Parallelisierung der Daten erfolgen. Das Hauptproblem waren dabei vor allem unterschiedliche bzw. fehlende Umlautkodierungen, die mithilfe verschiedener Algorithmen, Heuristiken und Tools (wieder-)hergestellt werden mußten.

(iii) Während bei den Nomina naturgemäß der Schwerpunkt auf der Morphologie lag, liegt bei der Kodierung der Verben das Gewicht eher auf syntaktischen Aspekten, wie z.B. Subkategorisierung etc. Ein Datenbankschema für Verben wurde entworfen, wobei die Unterstützung der Konsistenzprüfung der syntaktischen Kodierung der Valenzrahmen im Vordergrund stand. Bei der Erstellung der Daten waren offensichtlich inhaltliche Abhängigkeiten innerhalb eines Eintrags nicht überpüft worden. So setzt z.B. ein Korrelatspronomen die Existenz eines satzwertigen Komplements voraus, ohne daß dies durch die Kodierung sichergestellt war. In einem ersten Schritt wurde deshalb bei den Verben zunächst einmal die Konsistenz der Daten gemäß der originalen Kodieranweisung des SADAW-Lexikons überprüft und hergestellt. Darüberhinaus erfolgte ein erster Abgleich mit anderen Lexika im Sinne von (ii).

Die syntaktische Kodierung der Verbvalenz wurde danach einer gründlichen Restrukturierung unterzogen. Dies betraf zum einen das Problem einer effizienten Überprüfung und Rekodierung der alten SADAW-Daten, zum anderen die Entwicklung einer adäquaten Repräsentation in einem neuen Format. Das erste Problem ergab sich daraus, daß in den alten SADAW-Daten die Verbvalenz kumulativ repräsentiert war, d.h. zwei oder mehr mögliche verschiedene Konstruktionsrahmen wurden nicht disjunktiv aufgelistet oder als Alternativen gekennzeichnet, sondern zu einer Liste möglicher Komplemente zusammengeworfen. Um ein einfaches Beispiel zu nennen, wenn ein Verb entweder nur mit einem direkten Objekt *oder* nur mit einem Satzkomplement konstruiert werden konnte, so wurde dies in SADAW gleich kodiert, wie wenn ein Verb mit einem direkten Objekt *und* einem Satzkomplement stehen konnte.

Diese Information galt es nun, zu trennen und so zu repräsentieren, daß es gleichzeitig

auch möglich war, Optionalität adäquat darzustellen. Dabei zeigte sich, daß ohne eine geeignete semantische Lesartenunterscheidung, der Begriff der Optionalität keinen Sinn macht, da die Präsenz oder Absenz eines Komplements durchaus von einer Lesart abhängen kann. Die möglichen Valenzen wurden deshalb vollständig voneinander getrennt, indem die möglichen Kombinationen bei jedem Wort ausmultipliziert wurden, wobei Heuristiken entwickelt und angewandt wurden, durch die unmögliche Kombinationen ausgeschlossen wurden. Zur Korrektur der Daten wurde sodann eine ausgefeilte graphische Benutzeroberfläche geschaffen, durch die die falschen Valenzen manuell leicht und effizient von den korrekten getrennt werden können. Dieser Teil des Datenbanksystems ist ausführlich in der Diplomarbeit von Sabine Buchholz [Buc96] dokumentiert.

(iv) Zusätzlich wurde eine weitere Datensammlung aus dem SFB 100, das DEUSEM Wörterbuch mit semantischer Klassifikation von ca. 70.000 Nomina, für eine Wiederverwendung aufbereitet. Dazu wurde die Datensammlung zunächst auf Konsistenz der Kodierung hin untersucht und verbessert. Sodann wurde die ursprüngliche Kodierung in ein Sortensystem überführt, das als TDL-Verband interpretiert werden kann. Ein Problem ist allerdings, daß von diesen 70.000 Nomina sich nur um ca. 20.000 Einträge mit dem SARDICNomenlexikon überschneiden, so daß nur für diese Einträge auch eine vollständige syntaktische und morphologische Kodierung vorliegt. Entsprechende Erweiterungen für die anderen Daten sind geplant.

(v) Da sich die anfangs verwendete Datenbank FoxPro nach eine bestimmten Zeit als begrenzt geeignet bzw. teilweise veraltet herausstellte, wurde ein Wechsel der Softwarebasis ins Auge gefaßt und es wurden 12 neueste Programmierumgebungen und Softwarepakete ausführlich getestet und evaluiert. Die vorgegebenen Ansprüche und Evaluierungskriterien, als da waren Multiplattformfähigkeit, hohe Performanz und stabile und effiziente Programmierumgebung für Datenbank und Benutzerschnittstelle, erfüllte keines der Produkte zur vollen Zufriedenheit, wobei sich für MS Visual FoxPro 3.0 mit Erweiterungen (DLLs) in Visual C++ noch die besten Ergebnisse ergaben. Die Datenbank wurde deshalb entsprechend in der zweiten Hälfte des Projektes in Visual FoxPro rekodiert. Eine längerfrstige Trennung von Datenbankserver und Benutzerschnittstelle ist dabei vorbehalten.

(vi) Für die morphologischen Daten wurde eine Exportfunktion definiert, durch die aus den SARDIC-Daten ein Format erzeugt werden konnte, das mit der im PAGE-System verwendeten Morphologiekomponente kompatibel ist. Dadurch konnten erstmals größere Tests mit der verbesserten Morphologie gefahren werden. Aus den Daten wurde z.B. auch ein Vollformlexikon im Umfang von ca. 750.000 Formen generiert, das dem Bayrischen Archiv für Sprachdaten (BAS) in München zur Verfügung gestellt werden konnte. Das BAS vertreibt diese Daten unter der Bezeichnung PhonLex. Im Gegenzug wird das BAS eine phonologische Repräsentation dieser Formen liefern, die in das Lexikon integriert werden kann. Für die weitere Wartung der morphologischen Daten wurde ebenfalls eine benutzerfreundliche graphische Schnittstelle sowie die entsprechende Datenbank-Struktur entwickelt. Dieser Teil des Datenbanksystems ist ausführlich in der Diplomarbeit von Ingo Neis [Nei96] dokumentiert.

AP 24: Syntax

Neben einer Reihe von fortlaufenden Änderungen und Erweiterungen wurde das HPSG-Grammatikfragment des Deutschen vor allem in vier zentralen Punkten ausgebaut: Komplexe Sätze, Passiv und andere nicht-finite Konstruktionen, Verbvalenz und Optionalität, und Ge-

schlössen Wortklassen und Funktionswörter. • . . . . .

Zu den Komplexen Sätzen zählen sowohl subordinierte Komplementsätze als auch Relativsätze. Für die Analyse der Komplementsätze konnte auf die theoretischen Vorarbeiten für funktionale Köpfe und Verbstellung zurückgegriffen werden, wie sie z.B. in [Net96] beschrieben sind. Für Relativsätze wurde ein neuartiger Ansatz entwickelt, der im Gegensatz zum traditionellen HPSG-Ansatz eine Analyse ohne leere Knoten ermöglicht und gleichzeitig die Behandlung von Pied-Piping-Phänomenen erleichtert.

Des weiteren wurde das Fragment um Auxiliar- und Passivkonstruktionen ausgebaut. Damit steht im Bereich des Verbalkomplexes das volle Paradigma an synthetischen und analytischen Tempora zur Verfügung. Die Passivkonstruktionen umfassen momentan die Varianten des Vorgangs- und Zustandspassivs. Durch die Entwicklung eines Sortensystems zur Markierung von Vollverben, Tempus-Auxiliaren und Modalverben, konnten die relevanten kombinatorischen Restriktionen erfaßt werden, ohne daß auf Disjunktionen und Ambiguitäten zurückgegriffen werden mußte.

Es wurde eine neuartige Art der Behandlung von Valenzrahmen und Komplementation entworfen und implementiert. Anstelle der sonst in der HPSG üblichen Subkategorisierungslisten und entsprechenden Operationen werden Valenzen über zwei Merkmalsbündel repräsentiert, die eine Trennung der Spezifikation von Selektionsbeschränkungen und des Grades der Saturiertheit erlauben. Unter Ausnützung von sortaler Kodierung ist es dadurch möglich, das weitverbreitete Phänomen der Optionalität von Komplementen durch Unterspezifikation (anstelle von komplexen Disjunktionen) auszudrücken. Obzwar die Zahl der Regelinstanzen sich dadurch etwas erhöht, erhält man durch diesen Ansatz auch eine wesentliche Vereinfachung und Verbesserung der Behandlung von Wortstellungsvarianten.

Im Bereich der funktionalen Kategorien und geschlossenen Wortklassen wurden die (nichtkomplexen) präpositionalen Kategorien des Deutschen systematisch erfaßt und in TDL modelliert. Diese Klasse umfaßt vor allem alle Arten von Präpositionen, Präpositionaladverbien, Interrogativadverbien und nicht-derivative Formen von Adverbien. Sie wurde im Umfang von ca. 500 verschiedene lexikalischen Instanzen aufgearbeitet.

Durch die Überführung und Extraktion der morpho-syntaktischen Information des SarDic-Lexikons in eine MORPHIX-Repräsentation stehen der morphologischen Verarbeitungskomponente ca. 100.000 Nomen- und Adjektivstämme zur Verfügung. Für diese offenen Klassen wurden in der Grammatik Default-Einträge definiert, sodaß auch eine, wenn auch rudimentäre, syntaktische und semantische Abdeckung in diesem Umfang gegeben ist.

Durch die Kombination des PARADICE-Systems mit der TSNLP Testdatenbank ergaben sich Möglichkeiten für die Diagnose und Verbesserung der Grammatik, wie sie dieser Form vorher nicht erzielt werden konnten (s.u.). Ein Teil der Arbeit konzentrierte sich deshalb im wesentlichen auf die Bereiche, die aufgrund von systematischen Testläufen als fehler- und lückenhaft identifiziert werden konnten und die entsprechend bereinigt wurden.

Im wesentlichen sind dadurch alle im Antrag beschriebenen Phänomene und Aufgabenbereiche abgedeckt.

7.3 Verarbeitung

Der Teilbereich Verarbeitung konzentrierte sich auf die robuste und effiziente Verarbeitung von deklarativen Wissensquellen. Das Message-Extraction-System wurde hier im Rahmen der flachen Verarbeitungsstrategien entwickelt. In diesem Bereich war auch die Diagnose und Evaluierung des Kern-Systems und der Grammatik angesiedelt.

AP 31: Robustheit

Im Bereich "Robuste Verarbeitung" wurden die Spezifikation von Fehlermeldungsprotokollen definiert, die Grundlagen für die statistische Verarbeitung von Texten geschaffen und eine Methode zur robusten Behandlung unbekannter Wörter entwickelt.

Fehlermeldungsprotokolle: Fehlermeldungen wurden als Methoden für ein Objekt der Klasse ERROR definiert. Dieses Objekt verwaltet die Fehlermeldungen und selektiert entsprechende Fehlermeldungsroutinen. Der Vorteil der objekt-orientierten Modellierung ist, daß die Aktivierung von Fehlermeldungen und entsprechenden Routinen automatisch vom zugrundeliegenden OOP-System vorgenommen wird. Die entsprechende Klassendefinition wurde bereits in die Architektur übernommen.

Bei den Fehlermeldungsprotokollen werden "vorhersagbare Fehler" und "unvorhersagbare Fehler" unterschieden. Vorhersagbare Fehler sind solche, die vom Entwickler eines Moduls spezifiziert sind (gemäß eines vorgegebenen Standards). Solche Fehler sind "vorhersagbar", da der Modulzustand genau definiert und damit eine spezifische Fehlerbehandlung durchgeführt werden kann. Für die Definition von vorhersagbaren Fehlern wurde ein erster Prototyp implementiert und ebenfalls in die Architektur integriert. Unvorhersagbare Fehler sind solche, die nicht durch ein Modul als vorhersagbar definiert sind oder externe Ursachen haben (z.B. Hardwarefehler). In diesem Fall ist die Auswahl einer korrekten Fehlerbehandlung nicht eindeutig gegeben.

Korpusanalysetools: Die robuste Verarbeitung natürlicher Sprache (z.B. die Behandlung unbekannter Wörter) auf der Basis statistischer Information setzt Mechanismen zur automatischen Korpusanalyse voraus. Dazu wurde im Berichtszeitraum auf der Basis des existierenden Scanners und der Morphologiekomponente MORPHIX ein System implementiert, das für einen beliebigen in ASCII-Format vorliegenden Text, eine morphologisch basierte, automatische Korpusanalyse durchführt. Diese umfaßt eine lokale als auch eine globale Analyse.

In der lokalen Analyse wird für jede Satzeinheit folgende Information protokolliert: Identifikation von Tokens und Wortformen, morphologisch analysierte und nicht analysierte Formen, sowie die präterminale Kette eines Satzes in Form von Kategorien. Die globale Analyse protokolliert die absolute Häufigkeit von unbekannten Wörtern, von erkannten Wörtern, sowie von Clustern von präterminalen Ketten. Die pro Satz erkannten präterminalen Kategorienketten werden dabei in einem Diskriminationsnetz gespeichert, das später zum Definieren von korpusrelevanten endlichen Automaten herangezogen werden kann und damit als Grundlage für *Shallow Parsing* dienen kann. Des weiteren werden während der globalen Analyse n-grams für Wortformen, Lemmata und präterminale Elemente berechnet. Es ist an dieser Stelle zu betonen, daß die globale Analyse inkrementell arbeitet, also unmittelbar nach jeder lokalen Analyse einsetzt.

Das gesamte System ist bereits vielseitig parametrierbar. Beispielsweise kann definiert werden, welche der oben aufgeführte Informationen tatsächlich während der Analyse berücksichtig werden sollen. Auch die Berechnung der n-grams ist parametriersierbar (bi-grams,

tri-grams oder eben n-grams) und es kann angegeben werden, auf der Basis welcher Information (Wortform, Lemma oder präterminale Element) die n-grams berechnet werden sollen.

Dieses System wurde auf der Basis eines 700 KB großen Korpustextes (über 120.000 Tokens) getestet, wobei die gesamte Analysezeit bei ca. 200 CPU Sekunden liegt.

Robuste morphologische Verarbeitung: Ein Verbesserung bei der robusten Verarbeitung ergab sich vor allem im lexikalischen Bereich, wo gezielte Methoden für die Behandlung von inkorrektem und unbekanntem Input entwickelt wurden.

(i) Die Implementation der TRIEs zur effizienten Speicherung von MORPHIX Lexikoneinträgen wurde um den Lexikonzugriff mittels regulärer Ausdrücke und um eine Schnittstelle zu TDL erweitert. Bei den regulären Ausdrücken wird eine grep-ähnliche Syntax verwendet. So liefert z.B. der Ausdruck ".*ung" alle Einträge aus dem Lexikon, die das Suffix *-ung* besitzen, oder "n{i,a}cht" die Einträge für die Lemmata NICHT und NACHT. Zusätzlich zum regulären Ausdruck über den Wortstamm können beliebige Constraints spezifiziert werden, die die Ergebnismenge nochmals nach lexikonspezifischer Information filtern können. Dies erlaubt es z.B. alle im Plural umlautfähigen Nomen zu selegieren. Die Constraints können auf funktionaler Basis deklarativ und damit flexibel definiert werden.

Desweiteren wurde eine zusätzliche Operation FIND für die TRIEs implementiert. Diese Operation erlaubt es zu testen, ob für ein gegebenes Lemma eine spezifische Information (z.B. die morphologische Klasse) im Lexikon aufgeführt ist. Falls es sich bei dieser Information um eine TDL-Sorte handelt, kann der Zugriff auch mittels TDL-Typinferenz gesteuert werden.

Neben ihrem wichtigen Nutzen für allgemeine Lexikonverwaltungsaufgaben, bilden diese Erweiterungen die wesentlichen Grundlagen für die robuste Verarbeitung auf Wortebene, wie z.B. Rechtschreibkorrektur, das Erkennen und Tagging unbekannter Wörter, sowie einen semiautomatischem Lexikonerwerb. Es wurde damit begonnen, diese neuen Erweiterungen mit den Automatentools zu verzahnen (s.u.). Insbesondere wurde eine Methode zur Erkennung unbekannter Eigennamen entwickelt. Im Unterschied zu den einfachen schlüsselwortbasierten Methoden, werden hier insbesondere morpho-syntaktische und syntaktische Beschränkungen berücksichtigt. Auf diese Weise erkannte Eigennamen werden automatisch in ein temporäres Lexikon aufgenommen.

(ii) Eine verbesserte, robuste Verarbeitung wird auch bereits durch das Message Extraction System (s.u.) sichergestellt, das gezielt nur den "relevanten" Teil der Information in einem Text untersucht. Um jedoch auch eine zuverlässige Verarbeitung dieser relevanten Teile zu erlauben, wurde die Kompositabehandlung von MORPHIX (s.o.) um robuste Verfahren erweitert. Hierbei wird eine Analyse auch für den Fall geliefert, in dem nur Teile des Kompositums von der morphologischen Analyse abgedeckt sind. Ist beispielsweise in dem Kompositum *Hausboot* nur der rechte Teil *Boot* lexikalisch bekannt, kann trotzdem für das gesamte Wort eine korrekte morphologische Analyse berechnet werden. Damit wird ein zu frühes Scheitern der gesamten Analyse eines Satzes verhindert. Wie erste Performanzuntersuchungen ergeben haben, führt dieses Verfahren aufgrund des hervorragenden Effizienzverhaltens von MORPHIX zu keiner nennenswerten Erhöhung des Gesamtaufwandes der Verarbeitung.

AP 32: Effizienz

Im Bereich der Effizienten Verarbeitung lag der Schwerpunkt auf dem Ausbau und der Integration der Komponente für Explanation-Based Learning (EBL), der Konzeption eines chart-basierten top-down Generators und der Implementierung einer "Shallow Parsing"-

Komponente.

Explanation-Based Learning: Es wurde ein direktes Interface zwischen dem Parser und der EBL-Komponente implementiert, um zu ermöglichen, daß von EBL gefundene mögliche Ableitungen zusammen mit dem Parser benutzt werden. Der Parser extrahiert auf Anfrage aus einem Parsingergebnis die Ableitungsgeschichte derart, daß sie während der Lernphase von EBL an der entsprechenden Stelle des Diskriminationsnetzes abgespeichert werden kann. Wenn EBL in der Anwendungsphase eine gültige Kette (oder Teilketten) erkennt, werden diese Strukturen wieder an den Parser übergeben, der sie dazu benutzt, die abgespeicherte Derivation quasi-deterministisch nachzuvollziehen.

Aufbauend auf dieses Interface wurde in einem zweiten Schritt eine erste Version einer engen Verzahnung von Parsing und EBL auf phrasaler Ebene realisiert. Dadurch wurde es möglich in der Trainingsphase Ableitungsmuster für Phrasen zu extrahieren und in der Anwendungsphase entsprechend einzusetzen. Damit erhöht sich nicht nur die Performanz, sondern auch die Flexibilität des Gesamtsystems, das nicht mehr auf vollständige Satzmuster eingeschränkt ist. Im einzelnen wurden dabei die folgende Arbeiten durchgeführt:

- Trainingsphase:

  Alle möglichen Teilbäume eines analysierten Satzes werden extrahiert, abstrahiert und in einem Diskriminationsnetz abgelegt.

  Für jeden Teilbaum werden "Chunks" berechnet, d.h. Merkmalsstrukturen, die nur die Informationen des Wurzelknotens und die abstrahierte Terminal-Kette kodieren.

  Für das Diskriminationsnetz wurde ein graphischer Browser, sowie Operationen zum Sichern und Einlesen der durch EBL berechneten Strukturen implementiert.

- Anwendungsphase:

  Für eine lexikalisch analysierte Wortkette werden alle matchenden Chunks bestimmt und anschließend mit der lexikalischen Information unifiziert.

  Alle erfolgreich identifizierten Chunks werden dem Parser übergeben, der diese unmittelbar in seine Chart einfügt.

  Der Parser versucht nun diese Chunks zu kombinieren, wobei Regeln der Kompetenzgrammatik herangezogen werden können.

In der Anwendungsphase können unterschiedliche Strategien verwendet werden, die die Auswahl zwischen alternativen Chunks steuern. So gibt es einen "exhaustiven" Modus, der alle möglichen Chunks bestimmt und diese dem Parser weiterreicht. Der Parser erhält dadurch die Möglichkeit, mit Hilfe seines Agendamechanismus selbst die Auswahl zu bestimmen, wodurch ein "backtracking" in Chunks realisiert ist. Es ist aber auch möglich, daß EBL nur die Chunks mit größter Spannweite liefert. In diesem Fall, muß EBL selbst das Backtracking verwalten. Dieses durch EBL gesteuerte Backtracking wird in einer nächsten Version realisiert werden.

Die "phrasale" EBL-Methode kombiniert auch morphologische und lexikalische Entscheidungsbäume, wobei der morphologische Entscheidungsbaum als EBL-basierter Tagger fungiert. Es hat sich dabei gezeigt, daß diese *kaskadierte* EBL-Architektur eine erhebliche Steigerung des deterministischen Verhaltens der gesamten EBL-Analyse zur Folge hat und somit zur Steigerung der Effizienz erheblich beiträgt. Um auch im Bereich der Speicherkomplexität der EBL-Komponente nennenswerte Effizienzsteigerungen zu erlangen, ist die erste Version einer Verwaltung bezüglich Zugriffshäufigkeit und -Zeitpunkt implementiert worden. Damit

kann bereits jetzt die Expansion und Komprimierung von phrasalen Templates dynamisch kontrolliert werden.

Bei der Entwicklung der EBL Methode wurde besonderer Wert auf Grammatikunabhängigkeit gelegt. Dies wird mittels einer getypten Schnittstelle erreicht. Getestet wurde die EBL-Komponente mit folgenden Grammatiken: einer englischen HPSG-Grammatik, die am CSLI, Stanford entwickelt wurde, einer großen deutschen Grammatik im Stile des PATR-II-Formalismus, einer HPSG-basierten Grammatik für das Japanische und der in dem Projekt PARADICE entwickelten HPSG Grammatik. Um EBL für diese unterschiedlichen Grammatiken einsetzen zu können, mußte nur der Generalisierungsschritt in der Trainingsphase bei der Bestimmung der lexikalischen Supertypen angepaßt werden.

Schließlich wurden die EBL-Komponente um eine intelligente Speicherverwaltungskomponente erweitert und es wurde eine erste Version der EBL Methode für die Generierung implementiert, die auf der Basis der *Minimal Recursive Semantics* MRS aufbaut.

Generator: Die Konzeption einer Earley-basierten Grammatikgenerierung wurde abgeschlossen und in eine algorithmische Spezifikation gebracht. Die Vorteile dieser Generierungsmethode gegenüber der im DISCO-System verwendeten Variante der "semantic head driven" Generierung sind die folgenden:

- Durch die Verwendung einer dynamischen Selektionsfunktion ist eine stärkere datengesteuerte Kontrolle möglich. Darüberhinaus kann durch Verwendung von Präferenzen zusätzlich eine statistisch-orientierte Auswahl getroffen werden.
- Der neue Algorithmus verwendet eine Chart. Damit werden redundante Berechnungen vermieden, die bei Backtracking entstehen würden.
- Als generische Kontrolle wird ein Agendamechanimus verwendet, wie er auch in ähnlicher Weise bereits im Parser eingesetzt wird. Damit steht auch ein sehr flexibler Regelauswahlmechanismus zur Verfügung, der ebenfalls präferenzbasiert gesteuert werden kann.
- Das neue Verfahren ist aufgrund seiner Flexibilität auf eine größere Klasse von Grammatiken anwendbar. Weiterhin garantiert die Earley-basierte Kontrollogik, daß der neue Algorithmus auch für eine größere Klasse von Grammatiken terminiert.
- Der chart-basierte Algorithmus läßt sich auch prinzipiell leichter mit der EBL Methode verbinden, wobei ähnliche Effizienzgewinne zu erwarten sind, wie sie bereits für das Parsen erreicht wurden.

Shallow Parsing: Es wurde eine Komponente zur Verarbeitung von Automaten mit entsprechendem Kompiler für eine "Shallow Parsing"-Strategie implementiert. Diese Verarbeitungskomponente wird als Basismaschinerie für ein Message-Extraction-System verwendet werden. Neben Erweiterungen im Bereich von MORPHIX zählt hierzu vor allem eine Kompilation erweiterter regulärer Ausdrücke in einen bestehenden ATN-Formalismus. Dies wurde notwendig, da die prototypische Implementation eines solches Systems (für die Verarbeitung von Vortragsankündungen) gezeigt hat, daß die Definition der Templates und Phrasenerkenner direkt im ATN-Formalismus zu aufwendig ist.

Neben den üblichen regulären Operatoren ist es möglich, neue terminale Kanten zu definieren. Diese können als 'terminale Symbole' in den regulären Ausdrücken verwendet werden. Mit Hilfe dieser terminalen Kanten können beliebige Prozeduren in die regulären Aus-

drücke eingebunden werden, die zusätzlich noch auf lokalen Registern operieren dürfen. Damit können nun in einer deklarativen Sprache Transduktoren geschrieben werden, die als Vermittler zwischen den verschiedenen Verarbeitungsschichten eines Message-Extraction-Systems dienen können. Beispielsweise wurde eine Klasse von Automaten definiert, deren terminale Kanten auf MORPHIX-Ausgabe arbeiten, und die Nominalphrasen erkennen und gegebenenfalls in Merkmalsstrukturen übersetzen. Diese Automaten können in der COSMA-Domäne der Terminvereinbarungen eingesetzt werden. Im Bereich flache Verarbeitung wurde eine neue Methode implementiert, die die bidirektionale Verarbeitung von Finite-State-Grammatiken erlaubt.

Dieser Prototyp eines Message Extraction Systems wurde kontinuierlich weiterentwickelt. Insbesondere wurde die erste Version eines Phrasenkombinierers zur Templateerzeugung integriert. Die wesentliche Eigenschaft dieses Verfahrens ist, daß nun endliche Automaten direkt mit lexikalischen Elementen verankert werden können. Die effiziente und zielgerichtete Verarbeitung dieser lexikalisch verankerten Automaten wird mittels einer neu entwickelten bidirektionalen flachen Analysestrategie garantiert. Das Message Extraction System ist mittels einer HTML Schnittstelle mit dem weitverbreiteten Internet-Browser Netscape gekoppelt. Dies erlaubt es nicht nur, die Analyseergebnisse mittels Hypertext-Methoden im Dokument unmittelbar zugänglich zu machen, sondern eröffnet auch die Möglichkeit einer Integration des Message Extraction Systems in komplexe Internet-Dienste.

Generische Datenstrukturen: Ein neues Module für die Verarbeitung von Diskriminationsbäumen wurde implementiert, das als gemeinsame Datenstruktur für verschiedene Komponenten verwendet werden kann. Dadurch konnte modul-spezifischer Code ersetzt werden. So verwendet zum Beispiel das EBL-Modul das neue Werkzeug um phrasale Templates zu indizieren, wohingegen die Morphologiekomponente es für die Verarbeitung von Nominalkomposita einsetzt. Das Modul ist darüberhinaus mit einem effizienten Matcher für reguläre Diskriminationsbäume ausgestattet. In EBL wird dieser Matcher dazu verwendet um unbekannte Wörter zu analysieren, in der Morphologiekomponente wird er für Tagging eingesetzt.

Kompilation: In Zusammenarbeit mit Prof. Vijay-Shanker von der University of Delaware und Prof. Robert Kasper (Ohio State University) wurde eine Methode entwickelt und implementiert, mithilfe derer HPSG-Grammatiken in Tree Adjoining Grammars kompiliert werden können. Diese Kompilation hat vor allem den Vorteil, daß sie zu stark lexikalisierten Grammatiken führt, d.h. es gibt keine Strukturen der Grammatik, die nicht unmittelbar mit einem lexikalischen Element verbunden sind. Dadurch kann der Suchraum erheblich eingeschränkt werden, sodaß keine Regel mehr betrachtet werden muß, die nicht durch einen Lexikoneintrag (ein terminales Element im String) lizensiert ist.

Desweiteren wurde eine Methode zur Berechnung einer kontextfreien Approximation einer HPSG Grammatik entwickelt. Ausgehend von einer Regelmenge und Lexikon wurde eine endliche Domäne auftretender Merkmalsstrukturen mittels Regelanwendung und Ergebnisfilterung berechnet. Es waren verschiedene Optimimierungen notwendig, um die Kompilationszeit in vertretbaren Maßen zu halten (z.B. Subsumptionstests, effiziente Indizierung). Es stelle sich jedoch heraus, daß dieser Ansatz zu Grammatiken mit sehr großen Regelmengen führt, die ihrerseits Performanzproblemen aufwerfen. Als Resultat wird deshalb in der Schwerpunkt zukünftiger Forschung eher auf der Entwicklung von Methoden zur Optimierung der Regelmengen einer abgeleiteten Grammatik liegen.

AP 33: Subsprachen

Als Basis für die effiziente und robuste Behandlung von Texten in der Domäne der Terminvereinbarungen mithilfe des Subsprachen-Ansatzes, wurde ein Korpus von ca. 400 freien E-mails analysiert. Diese Texte wurden zunächst durch MORPHIX mit morpho-syntaktisch3r Typinformation annotiert. Ambiguitäten bei den Tags wurden danach zu einem großen Teil manuell aufgelöst und es wurden die Grenzen von Phrasen (speziell Nominal-, Präpositional- und Adverbialphrasen) markiert, die im Rahmen der Message Extraction relevant sein können. Darüberhinaus wurden in größerem Umfang Zeitungskorpora (Frankfurter Rundschau, TAZ) unter dem Gesichtspunkt von Personennamen, Titeln und Berufsbezeichnungen analysiert.

Auf der Basis dieser Korpusanalysen wurden entsprechend im Zusammenhang mit der Entwicklung des Message Extraction Systems dann automatenbasierte Subgrammatiken für komplexe Zeit-und Datumsangaben, Personennamen, sowie für Nominal- (NP) und Präpositionalphrasen (PP) entwickelt.

Die Zeit- und Datumsgrammatik deckt neben einfachen Ausdrücken (wie z.B. "am 21. Oktober (19.00 Uhr)"), auch modifizierte Angaben (wie z.B. "morgen (Mi, 11.5.) früh") und koordinierte Ausdrücke (wie z.B. "vom 19. (8.00 h) bis einschl. 21. Oktober (18:00 h)"). Die Personennamengrammatik umfaßt auch Titelangaben (wie z.B. "Prof. Dr. Albert Einstein"). Diese Grammatik verfügt über ein Lexikon von Eigennamen mit mehr als 40.000 Einträgen, getrennt nach Nachnamen und weiblichen und männlichen Vornamen. Darüber hinaus erlaubt diese Grammatik die Behandlung von unbekannten Namen durch Formulierung von Wildcards oder Platzhaltern in entsprechenden Automaten. Die NP- und PP-Grammatiken umfassen neben einfachen Beschreibungen ebenfalls koordinierte Strukturen. Darüber hinaus verfügen diese Grammatiken über Schnittstellen zur morphologischen Typbeschreibung der HPSG-Quellgrammatik. Die Grammatiken sind modular gehalten und können gleichermaßen zusammen, als auch isoliert voneinander eingesetzt werden. Für die Extraktion von Informationen aus Wirtschaftstexten wurden spezielle Finite-State-Grammatiken entwickelt mit denen Währungsangaben (incl. Integer- und Kardinal-Zahlen, sowie koordinierte Ausdrücke) und Firmennamen erkannt und analysiert werden können. Die Grammatik für Firmennamen ist auch in der Lage, Ausdrücke zu erkennen, die unbekannte Wörter enthalten, indem sie Wahrscheinlichkeitsbedingungen verwendet.

Um diese Subgrammatiken dem DFKI Projekt COSMA verfügbar zu machen, wurde das Message Extraktion System SMES integriert. Im besonderen wurden bidirektionale Verbautomaten implementiert, welche die Extraktion von Fragmenten aus der COSMA Domäne ermöglichen. Verbautomaten beschreiben generische Beschränkungen für die Subkategorisierung und speziellen Adjunkten. Sie sind lexikalisch an Verben gebunden, wodurch eine lexikongesteuerte Aktivierung stattfindet. Zum Testen der Modularität der Subgrammatiken wurde ebenfalls auf der Basis von SMES ein Prototyp zur Extraktion von Wirtschaftsinformationen implementiert, wobei die zu extrahierenden Templates Informationen über Firmenname, Umsatz, Tendenz und Zeitraum tragen.

AP 34: Messungen und Evaluierung

Im Bereich Evaluierung und Performanzmessung wurde in Synergie mit dem EU-Projekt TSNLP (Test Suites for Natural Language Processing, LRE 62-089) eine wesentliche Erweiterung der Funktionalität in PARADICE erzielt. Aufbauend auf die Vorarbeiten in DiTo und in

enger Zusammenarbeit mit PARADICE wurden in TSNLP die Methodologie und Technologie für die systematische Diagnose und Evaluierung von NLP-Systemen entwickelt; darüberhinaus entstanden Test Suites für das Deutsche, Englische und Französische, die in ihrem Umfang (rund 5000 sorgfältig ausgewählte grammatische wie ungrammatische *test items* je Sprache), der Abdeckung (zentrale syntaktische Phänomene) und dem Grad der linguistischen Annotationen weit über bestehende Test Suites hinausgehen. Um die Wartung, Erweiterung, Anpassung und Anwendung der Testdaten bestmöglichst zu unterstützen, sind die TSNLP Test Suites in einer relationalen Datenbank mit zwei parallelen Implementierungen (ANSI C, FoxPro) organisiert, welche am DFKI entwickelt wurden.

Für die systematische Performanzmessung und Diagnose in PARADICE wurde die TSNLP-Datenbank nahtlos in das Kernsystem integriert, indem

• die relevanten Testdaten identifiziert und angepaßt wurden;

• eine bidirektionale Schnittstelle zwischen Kernsystem und Datenbank geschaffen wurde (mit Hilfe der ANSI C Version und des Common-Lisp *foreign function interface);*

• die Datenbank um ein PARADICE *user & application profile* erweitert wurde, in dem anwendungsspezifische Daten (s.u.) gespeichert werden; und

• das Kernsystem um eine automatisierte Diagnosekomponente erweitert wurde, die ohne Benutzerinnenintervention (etwa über Nacht) eine Menge von Testdatensätzen aus der Datenbank liest, diese nacheinander verarbeitet, bei Bedarf die Verarbeitungsergebnisse mit den in der Datenbank spezifizierten erwarteten Resultaten abgleicht und eine Menge von Performanzparametern in die Datenbank zurückschreibt.

Die enge Verzahnung von PARADICE mit einem systematischen Diagnosewerkzeug breiter Abdeckung ist, soweit ersichtlich, einzigartig für eine moderne Grammatikentwicklungsumgebung.

Um die kontinuierliche Diagnose in der Grammatikentwicklung und den Vergleich zu früheren Verarbeitungsergebnissen (bezogen auf Abdeckung und Effizienz) zu ermöglichen, wurde die TSNLP-Datenbank um ein sogenanntes PARADICE *user & application profile* erweitert, so daß pro Testdatensatz anwendungsspezifische Informationen wie die Zahl der Lesarten, die Verarbeitungszeiten und das erwartete Resultat (als semantische Formel(n) in *NLL)* abgelegt werden können;

Zur Performanzmessung in PARADICE wurden bisher drei vollständige Testzyklen mit jeweils 3674 Testdatensätzen (rund 75 % der deutschen Test Suite; im wesentlichen alle Testdaten ausschließlich der Koordinationsphänomene) durchgeführt. Diese Testläufe ermöglichten es,

(1) die Schnittstelle zur TSNLP-Datenbank und die Testumgebung in PARADICE zu testen und Softwarefehler zu korrigieren;

(2) die Abdeckung und Präzision der Analyse zu bestimmen, systematische Fehler (lexikalische Lücken, fehlerhafte Regelanwendung, Übergenerierung, Softwarefehler usw.) zu diagnostizieren und zu korrigieren; und

(3) Referenzdaten für zukünftige Testläufe zu bestimmen.

Mehrere systematische Fehlerquellen wurden erkannt und behoben; der Wert der Test Suite als Meß- und Diagnosewerkzeug in der Grammatikentwicklung zeigt sich insbesondere darin, daß etliche der Fehlerquellen (darunter schwere Softwareausfälle) auch im langjährigen manuellen Testbetrieb nicht entdeckt wurden. Die folgende Tabelle faßt die Meßergebnisse, klassifiziert nach Abweichungen zwischen erwarteten und tatsächlichen Resultaten, zusammen. Durch die gezielte Diagnose konnte zwischen dem zweiten und dritten Testlauf eine beträchtliche Verbesserung der Abdeckung und Präzision erreicht werden; die verbleibenden

Abweichungen — insbesondere in der ersten Kategorie — resultieren vor allem aus fehlender lexikalischer Abdeckung und teilweise aus fehlerhaften Klassifizierungen in der Test Suite selber.

Abweichungen erwartetes vs. tatsächliches Resultat	**2. Zyklus**	**3. Zyklus**
grammatische Testdatensätze ohne Lesart	868	751
ungrammatische Testdatensätze mit Lesart(en)	235	47
unbegründete Ambiguitäten ($\geq$ 3 Lesarten)	27	19
schwerwiegende Softwarefehler	7	0
Gesamtverringerung der Abweichungen	**28 %**	

Insgesamt erwiesen sich sowohl die Abdeckung und Qualität der PARADICE-Grammatik (rund 78 % der Test Suite werden korrekt als grammatisch erkannt oder als ungrammatisch zurückgewiesen) als auch der verwendeten Testdaten (die Auswahl der Phänomene und systematische Variation der Testdatensätze ermöglichten die Identifikation und Diagnose mehrerer systematischer und bisher unerkannter Fehlerquellen) als recht gut. Weitere Arbeiten im Bereich Performanzmessung und -evaluierung werden sich auf die Erweiterung der Grammatik und Test Suite konzentrieren. Es ist darüberhinaus geplant, die Testdaten auch als Trainingsdaten für die Performanzforschung einzusetzen.

Gemeinsam mit TSNLP hat PARADICE die Test-Suite-basierten Evaluationsverfahren nicht nur weiter vorangetrieben, sondern es wurde auch ein Beitrag zur Standardisierung im Bereich Evaluation und Diagnose geleistet. So wurden die PARADICE-Evaluationswerkzeuge durch einen Subkontrakt zwischen Cray Systems (Luxembourg) und dem DFKI in die von der Europäischen Union entwickelte ALEP (Advanced Linguistic Engineering Platform) Referenzplattform integriert. Damit besteht erstmals die Möglichkeit zum plattformübergreifenden Vergleich von Diagnose- und Benchmark-Ergebnissen.

8 Voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse

Die Ergebnisse des PARADICE-Projects lassen sich sowohl in Form von unterschiedlichen Konfigurationen des Gesamtsystems als auch im Hinblick auf die einzelnen Module verwerten, in beider Hinsicht werden die Resultate des Projekts aktiv genutzt und weiterverwendet.

Als Gesamtsystem sind vor allem zwei Konfigurationen hervorzuheben, die PAGE Entwicklungsumgebung für Grammatiken basierend auf getypten Merkmalslogiken und das Message Extraction System SMES basierend u.a. auf Shallow Parsing Techniken.

Das PAGE-System wird mittlerweile zum einen im Verbmobil-II-Projekt am DFKI und am CSLI (Stanford) für die Entwicklung der deutschen, englischen und japanischen Grammatik im Rahmen des Teilprojekts 2 "Linguistische Analyse" eingesetzt. Wesentliche Teile dieses Systems werden in einer Weiterentwicklung auch in die Tiefe Verarbeitung in Verbmobil Eingang finden. Zum anderen wird PAGE inzwischen aber auch weltweit von anderen Institutionen verwendet, so z.B. von den amerikanischen Universitäten und Organisationen, die im English Resource Grammar Online (ERGO) Konsortium zusammengeschlossen sind. Dabei spielt unter anderem auch die enge Verbindung des PAGE-Systems mit der TSNLP-Datenbank eine wichtige Rolle.

Das Message Extraction System SMES bildete die Grundlage zweier weiterer Projekte. Zum einen wird die Weiterentwicklung des SMES-Systems in demPARADIME-Projektvorangetrieben. Zum anderen wurde in einem Industrieprojekt gezeigt, daß das System bereits in seinem jetzigen Stadium als Demonstrator-System problemlos auf neue Domänen angewendet werden kann. Die hohe Generalität und Flexibilität des Systems erwies sich vor allem darin, daß innerhalb kürzester Zeit und mit vergleichsweise geringem Aufwand drei verschiedene Typen von Anfragen innerhalb einer vollkommen neuen Domäne, der Online-Chronik der GECONIFOR Truppen in Bosnien, implementiert werden konnten.

Auf der Ebene der Module und Resourcen wären vor allem die Morphologie, das Lexikon und die Methoden zur Performanzverbesserung zu erwähnen.

Nachdem die Robustheit und Effizienz der Morphologiekomponente Morphix in PARADICE nochmals getestet und verbessert wurde, wird sie momentan in C++ portiert, um damit ein in sich geschlossenes, leicht portierbares Modul zu erhalten, das flexibel und modular auch in anderen Systemen und auf anderen Plattformen eingesetzt werden kann. Ein wichtige Rolle bei dieser Integration in andere Umgebungen spielt auch die systemübergreifende und theorieunabhängige morpho-syntaktische Schnittstelle, die eine schnelle Anbindung andere Grammatiken gewährleistet.

Mithilfe der Morphixkomponente konnte aus den SARDIC-Daten ein Vollformenlexikon des Deutschen erstellt werden, das ca. 750.000 Wortformen umfaßt. Dieses Vollformenlexikon wurde an das Bairische Archiv für Sprachdaten (BAS) geliefert und mit automatischen Phonologisierungskomponente zu einem Phonologisch-Phonetischen Lexikon erweitert. Dieses Lexikon wird unter anderem durch die European Linguistic Resource Agency (ELRA) in Paris unter dem Namen PhonLex verbreitet und kann als Basis sowohl für Text-to-Speech Komponenten als auch für Transkriptionssysteme dienen.

Das SARDIC-System (das Wörterbuch, die Datenbank und die Benutzerschnittstelle für Syntax und Morphologie) werden in anderen Projekten weiter ausgebaut und gepflegt. Dieses System stellt eine computer-lexikographische Resource dar, wie sie nach unserem Wissen für das Deutsche ansonsten nicht existiert. Sobald die Daten weiter bereinigt worden sind, wird

das System auch einer breiteren Gruppe von Benutzern zugänglich gemacht werden.

Ein weiteres wichtiges Ergebnis, für das die Einsatzfähigkeit unabhängig vom Gesamtsystem nachgewiesen werden konnte, ist die performanz-oriente Methode des Explanation-Based Learning (EBL). So konnte der Ansatz in verschiedenen Umgebungen, z.B. am CSLI (Standford), integriert und installiert werden und führte in allen Fällen zu substantiellen Performanzsteigerungen.

Die Ergebnisse des PARADICE-Projektes wurden regelmäßig auch auf der CeBit dem internationalen Fachpublikum vorgestellt und fand dort reges Interesse. Auf dem DFKI-Stand bei der CeBit 1995 wurde neben dem COSMA-System auch ein Prototyp des Message-Extraction-Systems für E-Mails präsentiert, das in der Domäne von Veranstaltungsankündigungen eingesetzt werden kann. Dieses System untersucht zunächst, ob es sich bei einer E-Mail um eine Vortragsankündigung handelt, und extrahiert gegebenfalls die relevante Information (Titel/Sprecher/Zeit/Raum). Obwohl das System als Prototyp in seiner Abdeckung noch eingeschränkt war, fand es großen Anklang bei den Standbesuchern.

Das PAGE-System wurde 1996 in Verbindung mit den (TSNLP) Evaluationswerkzeugen unter dem Titel "Qualitätssicherung in der Sprachverarbeitung" ausgestellt. Das große Besucherinteresse zeigte, daß dem Bereich der Qualitätssicherung eine stetig wachsende Bedeutung zukommt.

Bei der CeBit 1997 konnte das SMES Message Extraction System angewandt auf die Domäne der Wirtschaftsmeldungen vorgeführt werden. Diese Variante des Systems kann interaktiv über einen WWW-Browser angesteuert werden, sodaß das Internet und Online-Potential des Systems unterstrichen wurde.

9 Fortschritte bei anderen Stellen

Fortschritte bei anderen Stellen waren vor allem bei den konkurrierenden Kernmaschinen festzustellen, die an anderen Stellen entwickelt wurden. Hier wäre zunächst die Weiterentwicklung des ALEP-Systems zu nennen, das weiter von der CEC gefördert wurde. Bei diesem System wurden vor allem Erfolge hinsichtlich der Effizienzsteigerung bei der Verarbeitung und bei der Entwicklung einer Methodologie zum Grammar Engineering erzielt.

Gleichfalls substantiell weiterentwickelt wurde die Core Language Engine (CLE) des SRI Cambridge. Diese System fand Eingang in eine Reihe von Anwendungen und wird in seiner neuesten Entwicklung auch erfolgreich in einem System zur Übersetzung gesprochener Sprache (Spoken Language Translation - SLT) eingesetzt.

Im Bereich der HPSG-orientierten Systeme wurde an der Universität Tübingen ein System namens TROLL entwickelt, das versucht, HPSG in einer möglichst reinen Form zu verarbeiten und das auch Frotschritte im Hinblick auf die Verarbeitung lexikalischer Regeln erzielte. Das ALE-System von Bob Carpenter wurde zwar weiter entwickelt, stellt aber keine vergleichbare Konkurrenz zum PAGE-System dar.

Während der Laufzeit des Projektes wurde ein Antrag auf einen neuen interdisziplinären SFB für *Ressourcenadaptive kognitive Prozesse* an der Universität des Saarlandes von der DFG positiv begutachtet, und es wurden von 12 beantragten Projekten 11 Projekte mit einer Gesamtausstattung von ca. 25 Mitarbeitern genehmigt. An diesem SFB sind neben Lehrstühlen in der Psychologie, Computerlinguistik, Informatik und Philosophie, auch die vier Fachbereichsleiter am DFKI Saarbrücken mit Projekten beteiligt. Der Bereich Computerlinguistik (Prof. Uszkoreit) ist im wesentlichen mit zwei Projekten vertreten, die beide auf effiziente und robuste Verarbeitung abzielen.

10 Erfolgte oder geplante Veröffentlichung der Ergebnisse

Literatur

[Buc96] BUCHHOLZ, S.: *Entwicklung einer lexikographischen Datenbank für die Verben des Deutschen.* MA thesis, Universität des Saarlandes, Saarbrücken, 1996.

[CGN] CROUCH, R., GAIZAUSKAS, R. und NETTER, K.: *Report of the Study Group on Assessment and Evaluation.* Technical Report, Cambridge, Sheffield, Saarbrücken, 1995.

[DFK+94] DALE R., FINKLER W., KITTREDGE R., LENKE N., NEUMANN G., PETERS C., und STEDE M.: *Report from Working Group 2: Lexicalization and Architecture.* In: HOEPPNER W., HORACEK H. und MOORE J., (Herausgeber): *Principles of Natural Language Generation* . Schloß Dagstuhl, Saarland, Germany, 1994.

[DEM+96] DÖRRE, J., ERBACH, G., MANANDHAR, S., SKUT, W. und USZKOREIT, H.: *A Report on the Draft EAGLES Encoding Standard for HPSG.* In: *International HPSG Conference,* Marseille, 1996.

[EM95] ERBACH, G. und MANANDHAR, S.: *Visions for Logic-Based Natural Language Processing.* In: *Proceedings of the ILPS '95 workshop "Visions for the Future of Logic Programming",* Portland, Oregon, 1995.

[Erb95] ERBACH, G.: *Why NLP needs Oz.* In: *Proceedings of the First International Workshop on Oz Programming,* Martigny, Switzerland, 1995.

[Erb96] ERBACH, G.: *Bottom-Up Earley Deduction for Preference-Driven Natural Language Processing.* PhD thesis, Universität des Saarlandes, Saarbrücken, 1996.

[EvdKM+95] ERBACH, G., KRAAN, M. v. D, MANANDHAR, S., RUESSINK, H., SKUT, W. und THIERSCH, C.: *Extending Unification Formalisms.* In: *2nd Language Engineering Convention,* London, 1995.

[KK96] KASPER, W. und KRIEGER, H.-U.: *Modularizing Codescriptive Grammars for Efficient Parsing.* In: *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96,* 1996.

[KKNVS95] KASPER, R., KIEFER, B., NETTER, K. und VIJAY-SHANKER, K.: *Compilation of HPSG to TAG.* In: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Seiten 92-99, Cambridge, Mass., 1995.

[KS94a] KRIEGER, H.-U. und SCHÄFER, U.: *TDL—A Type Description Language for Constraint-Based Grammars.* In: *Proceedings of the 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan,* Seiten 893-899, 1994.

[KS94b] KRIEGER, H.-U. und SCHÄFER, U.: *TDL— A Type Description Language for HPSG. Part 1: Overview.* Research Report RR-94-37, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany, 1994.

[KS94c] KRIEGER, H.-U. und SCHÄFER, U.: *TDL— A Type Description Language for HPSG. Part 2: User Guide.* Document D-94-14, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany, 1994.

[Kri95] KRIEGER, H.-U.: *Typed Feature Formalisms as a Common Basis for Linguistic Specification.* In: STEFFENS, P. (Herausgeber): *Machine Translation and the Lexicon.* Springer, Berlin, 1995.

[LORP+96] LEHMANN, S., OEPEN, S., REGNIER-PROST, S., NETTER, K., LUX, V., KLEIN, J., FALKEDAL, K., FOUVRY, F., ESTIVAL, D., DAUPHIN, E., COMPAGNION, H., BAUR, J., BALKAN, L. und ARNOLD, D. TSNLP — Test Suites for Natural Language Processing. In *Proceedings of COLING 1996*, pages 711 - 716, Kopenhagen, 1996.

[Nei96] NEIS, I.: *SarDic: Das morphologische Lexikon und der morphologische Server.* Dipl.-Arbeit, Universität des Saarlandes, Saarbrücken, 1996.

[NNP94] NERBONNE, J., NETTER, K. und POLLARD, C. (Herausgeber): *German in Head-Driven Phrase Structure Grammar.* Nummer46 in *CSLI Lecture Notes.* CSLI, Stanford, 1994.

[NKKVS94] NETTER, K., KASPER, R., KIEFER, B. und VIJAY-SHANKER, K.: *HPSG and TAG.* In: *In: 3^{e} Colloque International sur les Grammaires d'Arbres Adjoints. (TAG+3). Rapport Technique TALANA-RT-94-01*, Seiten 77-82, Paris, 1994.

[Net94a] NETTER, K.: *Syntax in der Maschinellen Sprachverarbeitung.* Informationstechnik und Technische Informatik, 36(2):6-13, 1994.

[Net94b] NETTER, K.: *Towards a Theory of Functional Heads: German Nominal Phrases.* In: NERBONNE, JOHN et al. [NNP94], Seiten 236-280.

[Net96] NETTER, K.: *Functional Categories in an HPSG for German.* Doktorarbeit, Universität des Saarlandes, Saarbrücken, 1996.

[Neu94a] NEUMANN, G.: *A Uniform Computational Model for Natural Language Parsing and Generation.* Doktorarbeit, Universität des Saarlandes, Germany, 1994.

[Neu94b] NEUMANN, G.: Application of Explanation-based Learning for Efficient Processing of Constraint-based Grammars. In *Proceedings of the Tenth IEEE Conference on Artifical Intelligence for Applications*, pages 208 - 215, San Antonio, Texas, 1994.

[Neu96] NEUMANN, G.: *Interleaving Natural Language Parsing and Generation Through Uniform Processin.* Research Report RR-96-03, DFKI GmbH, Saarbrücken, March 1996. submitted to AI Journal.

[NvN94] NEUMANN, G. UND NOORD, G.-J. VAN: *Reversibility and Self-Monitoring in Natural Language Generation.* In: STRZALKOWSKI, T., (Herausgeber): *Reversible Grammar in Natural Language Processing*, pages 59 - 96, Kluwer, 1994.

[NvN96] NOORD, G.-J. VAN UND NEUMANN, G.: *Syntactic Generation.* In: COLB, R., MARIANI, J., USZKOREIT, H., ZAENEN, A. und ZUE, V. (Herausgeber): *Survey of the State of the Art in Human Language Technology*, Kapitel 4. Cambridge University Press and Giardini, 1996. In press.

[ONK96] OEPEN, S., NETTER, K., und KLEIN, J. TSNLP — Test Suites for Natural Language Processing. In John Nerbonne, editor, *Linguistic Databases*, CSLI Lecture Notes. Center for the Study of Language and Information, 1997. forthcoming.

[UBC+96] H. USZKOREIT, R. BACKOFEN, C. CALDER, C. CAPSTICK, L. DINI, J. DÖRRE, G. ERBACH, D. ESTIVAL, S. MANANDHAR, A.-M. MINEUR, und S. OEPEN. The EAGLES formalisms working group — final report. Technical report, Expert Advisory Group on Language Engineering Standards (LRE 61-100), 1996.