Abschlussbericht

COLLATE-UdS

Computational Linguistics and Language Technology for Real-World Applications

Universität des Saarlandes
FR 4.7 Computerlinguistik
und
FR 6.7 Informatik
66041 Saarbrücken

Prof. Dr. Manfred Pinkal
Prof. Dr. Hans Uszkoreit
Prof. Dr. Wolfgang Wahlster
Dr. Gregor Erbach
Christian Braun
Gerhard Fliedner
Christian Müller
Dr. C. J. Rupp
Rainer Wasinger
Tianfang Yao
Zhiping Zheng

Juni 2004

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IN A01 B gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

I Kurzdarstellung des Projekts

I.1 Aufgabenstellung

Gegenstand des Vorhabens ist die Vorlauf- und Anschubforschung eines nationalen Kompetenzzentrums für Sprachtechnologien, die noch nicht am Markt vorhanden sind. Die Kompetenz soll der F&E-Gemeinschaft, den deutschen Firmen, die sprachtechnologische Produkte entwickeln und vermarkten, sowie potenziellen Anwendern zugute kommen. Sie soll aber auch Saarbrücken und damit Deutschland im internationalen Wettbewerb als herausragenden F&E-Standort der Sprachtechnologie stärken.

Das Projekt COLLATE dient der Stärkung der internationalen Stellung der deutschen F&E in dem strategisch wichtigen Bereich der Sprachtechnologie. Sprachtechnologie ist eine wesentliche Basiskomponente für zukünftige Informations- und Kommunikationstechnologien. Ein besonderer Schwerpunkt des Projekts ist der Brückenschlag von aktuellen Sprachtechnologieforschung hin zur industriellen Verwertung. Das Hauptziel ist dabei die Beschleunigung des Transfers aus der Forschung in praktische Anwendungen. Zur Erreichung dieses umfassenden und anspruchsvollen Ziels besteht das Projekt aus mehreren komplementären, interagierenden Komponenten.

Ein zentraler Bestandteil des Anschubvorhabens sind intensive Forschungs- und Entwicklungsarbeiten in drei Kern-Technologiebereichen der Sprachtechnologie:

- Sprachbasierte Informationsextraktion und -fusion
- Dialogische Interaktion für Wissenszugang und -erwerb
- Sprachbasiertes Informationsmanagement und -retrieval

Diese F&E-Arbeiten wurden an der Universität der Saarlandes durchgeführt, um theoretische und praktische Ergebnisse aus den zahlreichen auf langfristige Grundlagenforschung angelegten Projekten der Arbeitsgruppen der Antragsteller, sowie anderer einschlägiger Forschungsvorhaben des SFB 378, des Graduiertenkollegs "Sprachtechnologie und Kognitive Systeme" und der Fachrichtungen Computerlinguistik und Informatik aufzunehmen und in Technologieverbesserungen umzusetzen.

I.2 Voraussetzungen der Vorhabensdurchführung

Die Sprachtechnologie wird heute unbestritten als eine Schlüsseltechnologie für die Zukunft der IT-Industrie, für den Ausbau der Informationsgesellschaft und für die technologische Infrastruktur der Wissensgesellschaft gesehen. Die Sprachtechnologie ist der Sammelbegriff für eine ganze Klasse von einzelnen Technologien, die durch algorithmisiertes Wissen über die Eigenschaften menschlicher Sprachen auf den Umgang mit Texten oder gesprochenen Äußerungen spezialisiert sind. Sie ermöglichen eine Fülle von Anwendungen, die von Diktiersoftware bis zu Dialogsystemen, von der Indizierung und Navigation auf dem WWW bis zur maschinellen Übersetzung reichen. Die wissenschaftlichen Grundlagen dieser Technologie kommen zu einem großen Teil aus der Computerlinguistik, aber zu geringeren Teilen auch direkt aus der Informatik, den Sprachwissenschaften und der Akustik.

Es gibt weltweit einige Standorte, an denen gezielt versucht wird, die notwendige Breite in der Sprachtechnologie herzustellen. Das ist aber erst ansatzweise erfolgreich. In Pittsburgh (USA) an der CMU sind zwar die Sprachtechnologien in bemerkenswerter Breite vertreten, es fehlen aber die Sprachwissenschaften und die computerlinguistische Basisforschung. In Stanford (USA) hingegen ist die Grundlagenforschung in Linguistik und Informatik hervorragend ausgebaut, es mangelt aber an Sprachtechnologien, die man höchstens in einigen Industriefirmen in der Umgebung findet. Nur an der Universität von Edinburgh (UK), wo sich in der computer-

linguistischen Grundlagenforschung der Bogen von den Kognitionswissenschaften bis hin zur theoretischen Informatik spannt, sind zumindest einige der wichtigsten Sprachtechnologien auch an der Universität vertreten. Deutschland kann zwar hervorragende Computerlinguisten und Sprachtechnologen aufweisen, nur sind diese über die Republik verstreut. Das bisher größte deutsche Forschungsvorhaben der Sprachtechnologie, der BMBF-Verbundprojekt Verbmobil, brauchte etwa zwanzig deutsche Partner, um alle relevanten Aspekte der Sprachverarbeitung gut abzudecken. Eine gewisse Breite findet sich lediglich in Saarbrücken, dem bedeutendsten deutschen Zentrum der Computerlinguistik und Sprachtechnologie, wo auch die Koordination des Verbmobil-Vorhabens angesiedelt war und derzeit auch die Projektleitung des MTI-Leitvorhabens SmartKom sitzt.

An den wenigen Standorten, die dem Idealbild des breit ausgestatteten Kompetenzzentrums nahe kommen, sind jeweils mehrere neue Sprachtechnologiefirmen entstanden, so z.B. in Pittsburgh, der Stanford-Umgebung, Edinburgh und Saarbrücken. Im Gegensatz zu den reiferen Ingenieurwissenschaften haben sich aber bisher in der Sprachtechnologie noch keine Zentren herausgebildet, die durch die umfassende Abdeckung der Teilgebiete und die enge Verbindung von wissenschaftlicher Grundlagenforschung, der Weiterentwicklung der Basistechnologien und Entwicklung realistischer Anwendungen zum Motor der internationalen technologischen Fortschritts und zum Werkzeug der Beschleunigung des Technologietransfers werden konnten.

Wie ständige Anfragen klar gezeigt haben, gibt es für ein solches Zentrum einen dringenden Bedarf. Durch das Bereitstellen von F&E-Kompetenz ist ein solches Zentrum ein wichtiger Mittler zwischen der Anwendergemeinschaft und den Technologieentwicklern.

I.3 Planung und Ablauf des Vorhabens

Das Vorhaben COLLATE wurde am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) und an der Universität des Saarlandes durchgeführt. Das Gesamtvorhaben wurde vom DFKI koordiniert. Am DFKI waren die Forschungsbereiche Sprachtechnologie (Prof. Uszkoreit) und Intelligente Benutzerschnittstellen (Prof. Wahlster) beteiligt. An der Universität des Saarlandes waren die Fachbereiche Computerlinguistik und Informatik beteiligt. Die Forschung wurde in drei Arbeitsgruppen durchgeführt, die von Prof. Pinkal (Computerlinguistik), Prof. Uszkoreit (Computerlinguistik) und Prof. Wahlster (Informatik) geleitet wurden. Das Projekt ist in vier Arbeitspakete gegliedert:

AP 1: Sprachbasierte Informationsextraktion und –fusion

Die Ergebnisse sind dieses Arbeitspakets sind in Abschnitt II.1.1 ausführlich dargestellt.

AP 2: Dialogische Interaktion für Wissenszugang und –erwerb

Die Ergebnisse dieses Arbeitspakets sind den Abschnitten II.1.2.3 (Dialogmodellierung für Informationszugriff) und II.1.5 (Multimodale Dialogarchitektur und Benutzergruppenerkennung) ausführlich dargestellt.

AP 3: Sprachbasiertes Informationsmanagement und –retrieval

Die Ergebnisse dieses Arbeitspakets sind den Abschnitten II.1.2 (NLP-Framework für Informationsmanagement), II.1.4 (Web-Korpora), und II.1.3 (Question Answering) ausführlich dargestellt.

AP4: Technologie-, Infrastruktur- und Projektkoordination

In diesem Arbeitspaket wurde die Zusammenarbeit mit dem DFKI und mit externen Partnern koordiniert, und der Austausch von Ressourcen und Technologien durchgeführt. Außerdem wurde die technische Infrastruktur (Linux-Cluster, Dialoglabor, Usability-Labor, Tonstudio, Gestenerkennungslabor) spezifiziert, beschafft, installiert und in Betrieb genommen.

Ein wesentliches F&E-Ergebnis des Projekts ist eine umfassende Informationsextraktions-Plattform mit einem Grammatikformalismus, Entwicklungswerkzeugen, Laufzeitsystem, Evaluationswerkzeugen, multilingualen Ressourcen, multilingualen Grammatiken, sowie Entwicklungs- und Testdaten. Forschungsergebnisse im Bereich von NLP-Methoden für Informationsmanagement sind ein Framework mit neuen und robusten Methoden für Chunking und topologisches Parsing, semantische Analyse und Annotation auf Basis von FrameNet und die praktische Anwendung eines Informationszustandsmodells (information state approach) für Dialogmanagement. Im Bereich von Question-Answering-Systemen für offene Domänen wurden erhebliche Verbesserungen in Bezug auf Geschwindigkeit, Robustheit und Skalierbarkeit eines existierenden Question-Answering-Systems erzielt. Die Wissensbasis des Systems wurde um acht Jahrgänge von CNN-Nachrichten erweitert. Wir haben Methoden und Werkzeuge entwickelt, um Web-Korpora mit umfangreichen Metadaten über Dokumente und Hyperlinks zu akquirieren und zu verwalten. Forschungsergebnisse im Bereich von mobilen, multimodalen und modularen Schnittstellen sind eine fortgeschrittene Systemarchitektur mit einer Medienfusionskomponente und effektive Algorithmen für die Erkennung von Benutzergruppen verschiedenen Alters und Geschlechts, anhand von akustischen Stimmmerkmalen.

Informationsextraktion

Die Anzahl der weltweit verfügbaren Informationen wächst ständig, so dass die Auswertung der Informationsquellen ohne leistungsfähige Werkzeuge nicht mehr machbar ist. Im Projekt COLLATE wurde ein System zur gezielten Extraktion von Informationen aus großen Textmengen entwickelt, das für neun europäische und asiatische Sprachen verfügbar ist. Mit dem System SProUT können gezielt Informationen über bestimmte Personen, Firmen oder Ereignisse aus Texten extrahiert und in eine Datenbank eingetragen werden.

SProUT wird eingesetzt zur automatischen Auswertung von Reisewarnungen für Krisengebiete, und für die Extraktion von Kundenmeinungen über Elektronik-Geräte und Kraftfahrzeuge aus Online-Diskussionsforen.

Question-Answering

Sucht ein Benutzer eine Antwort auf eine gezielte Frage, so muss er mit heutigen Suchmaschinen Stichwörter für eine Anfrage aussuchen, und dann mit großem Aufwand suchen, ob eines der gefundenen Dokumente die Antwort auf seine Frage enthält. In COLLATE wurde ein System weiterentwickelt und optimiert, das diese Schritte überflüssig macht, und direkt auf eine natürlichsprachliche Frage einen Satz mit der passenden Antwort liefert. Das System AnswerBus analysiert dazu die Frage des Benutzers, erstellt Stichwörter für eine Anfrage an eine Suchmaschine, und extrahiert Antworten aus den Ergebnisseiten. Auch das Abfragen von Nachrichtenquellen ist möglich. Das System wird täglich von einigen tausend Benutzern verwendet. Es wurde eine gesprochene Eingabe mit Spracherkennung realisiert, so dass demnächst Fragen auch telefonisch beantwortet werden können.

Mobile Assistenten für alle Benutzergruppen

Mit dem System M3I wurde eine Anwendung entwickelt, die Informations- und Navigationssysteme auf mobilen Geräten wie Handys oder PDAs verfügbar macht. Um eine komfortable Interaktion zu ermöglichen, wurden die Geräte mit Sprachein- und -ausgabe ausgestattet. Außerdem können sprachliche Eingaben mit Zeigegesten verbunden werden. Damit kann der Benutzer beispielsweise auf seinem Display auf einen Punkt auf dem Stadtplan zeigen und fragen „Wie komme ich dorthin?“ oder „Was ist das für ein Gebäude?“. Um die Geräte für möglichst viele Benutzergruppen verwendbar zu machen, optimiert sich die Spracherkennung automatisch für Geschlecht und Altersgruppe des Benutzers.

Internationaler wissenschaftlicher Beirat

Zur Sicherung der wissenschaftlich-technischen Qualität wurde ein wissenschaftlicher Beirat berufen, dem die folgenden international anerkannten Wissenschaftler angehören:

- Steven Bird, University of Melbourne und Linguistic Data Consortium
- Martin Kay, Stanford Univ. und XEROX PARC
- Sharon Oviatt, Oregon Graduate Institute of Science and Technology
- Donia Scott, University of Brighton
- Oliviero Stock, IRST Trento
- Hans Tillmann, BAS, Univ. München

Am 25. Oktober 2001 fand eine Sitzung des wissenschaftlichen Beirats statt. Dabei wurden die bisherigen Ergebnisse und der Arbeitsplan des Projekts vorgestellt und diskutiert. Die Empfehlungen des wissenschaftlichen Beirats wurden bei der weiteren Planung berücksichtigt. Ebenfalls am 25. Oktober 2001 fand die offizielle Eröffnung des Projekts statt, bei der hochrangige Vertreter aus Wissenschaft, Wirtschaft und Politik anwesend waren. Dabei wurden die Ziele und wissenschaftlichen Ansätze des Projekts präsentiert.

Am 7. Februar 2003 fand die Sitzung des wissenschaftlichen Beirats (Dr. Oliviero Stock, Prof. Dr. Hans Tillmann, Prof. Donia Scott, Prof. Dr. Martin Kay) statt. Dabei wurden Ergebnisse und Arbeitspläne vorgestellt. Von Seiten des wissenschaftlichen Beirats wurde die hohe Qualität der Projektarbeit und die gelungene Verbindung von innovativen Anwendungen und theoretischer Grundlagenarbeit gelobt. Das Projekt wurde als Modell für größere Aktivitäten, z.B. im europäischen Rahmen, bezeichnet.

I.4 Wissenschaftlicher und technischer Ausgangsstand

Das Ziel des Informationsmanagements besteht darin, das Informationsbedürfnis von Organisationen zu erfüllen. Die praktische Aufgabe ist daher die Verwaltung und Nutzbarmachung von sehr großen Informationsmengen. Diese soll dem einzigen Zweck dienen, die Information den Entscheidungsträgern genau dann zu liefern, wenn sie benötigt wird, und sie überdies so zu präzisieren, dass sie den Nutzer bei seiner Entscheidung effizient und effektiv unterstützt. Im folgenden wird der wissenschaftlich-technische Ausgangsstand in Einzelbereichen des natürlichsprachlichen Informationsmanagements dargestellt.

Informationsextraktion

Im Bereich der Informationsextraktion gab es vor allem Aktivitäten im englischsprachigen Bereich, wobei die technologische Entwicklung durch vergleichende Technologieevaluation im Rahmen der Message Understanding Conference (MUC) vorangetrieben wurde. Vergleichbare Evaluationen und Testdaten für die deutsche und andere europäische Sprachen waren nicht verfügbar.

In Saarbrücken gab es aus früheren Projekten eine Vielfalt von Forschungssystemen (SMES, MESON; SPPC), Ressourcen und Testdaten. Es fehlte jedoch eine benutzerfreundliche Entwicklungsplattform mit einem ausdrucksmächtigen Grammatikformalismus, Evaluationswerkzeug, sowie ein robustes, effizientes Laufzeitsystem.

Es gab nur Informationsextraktionssysteme für einzelne Sprachen, insbesondere für das Englische, jedoch noch keine Methodologie zur Entwicklung von parallelen Grammatiken für mehrere Sprachen, die eine systematische Wiederverwendung von Ressourcen (Ortsnamen, Personen- und Firmennamen) und Regelsystemen unterstützt, und damit die effiziente Entwicklung von Grammatiken für neue Sprachen ermöglicht.

Question Answering

Open-domain Question Answering (QA) hat sich seit einigen Jahren als wichtige Forschungsrichtung etabliert, die Techniken des Information Retrieval, der Sprachverarbeitung und auch einfache Wissensrepräsentation einsetzt, um Antworten auf Fragen aus großen Textmengen zu extrahieren. Diese Forschungsrichtung wurde und wird im Rahmen der Text Retrieval Conference (TREC) durch eine eigene Evaluation unterstützt. Dabei werden jedes Jahr anspruchsvollere Aufgaben gestellt, die vorab in einer Roadmap festgelegt wurden. Eine neue Forschungsrichtung waren web-basierte Question-Answering-Systeme. Zur Zeit der Antragstellung befand sich das web-basierte QA-System AnswerBus, das heute täglich von hunderten Benutzern eingesetzt wird, noch in Entwicklung[1].

Multimodalität und Benutzermodellierung

Es gab Systeme, die verschiedene Ein- und Ausgabemodalitäten als Alternativen unterstützten, jedoch erst erste Ansätze zur Medienfusion. Insbesondere gab es noch keine effiziente Medienfusionskomponente für Mobilgeräte wie PDAs.

Im Bereich der Benutzermodellierung gab es verschiedene Ansätze, um Systeme für bestimmte Benutzergruppen zu adaptieren, wobei die Zuordnung eines Benutzers zu einer bestimmten Gruppe vorausgesetzt wurde. Diese Voraussetzung ist gegeben, wenn registrierte Benutzer durch Benutzername und Passwort oder biometrische Methoden identifiziert werden. Forschungsbedarf bestand bei der Zuordnung nicht registrierter Benutzer zu Benutzergruppen mit bestimmten Eigenschaften wie Alter oder Geschlecht in offenen Systemen, z.B. aufgrund von Eigenschaften des Sprachsignals.

Dialog

In früheren Projekten wie TRINDI und SIRIDUS sind theoretisch fundierte Dialogmodelle erarbeitet worden, die auf einem Informationszustandsmodell beruhen. Praktisch angewandt wurden sie nur für aufgabenorientierte Dialoge, wie z.B. Reisebuchungen, aber noch nicht für die Verbesserung von Informationszugriff.

I.5 Zusammenarbeit mit anderen Stellen

Es gab enge Kooperation in der Forschung mit anderen Universitäten und Forschungsinstituten, ebenso wie mit Industriefirmen mit dem Ziel des Technologietransfers.

Besonders intensive Kooperationen gab es mit dem Deutschen Forschungszentrum für Künstliche Intelligenz, dem Koordinator des Verbundprojekts COLLATE. Das Informationsextraktionssystem SProUT wurde gemeinsam mit DFKI-Mitarbeitern entwickelt und evaluiert.

Auf verschiedenen Gebieten gab es Zusammenarbeit mit externen Experten. Auf dem Gebiet der gesprochenen Sprache gab es folgende Kooperationen:

Alan Black, CMU: Diskussionen über mögliche Portierung des SPHINX Recognizers auf Festkommaoperationen, um eine bessere Performance auf dem Strongarm Prozessor zu erreichen

Eric Woudenberg, Speechworks: Technische Modifikationen am iPAQ zur Verbesserung der Spracherkennung

Derek Jacoby, Microsoft Research: Diskussion über Distributed Speech Recognition

Dr. Francis Ganong, ScanSoft (Forschungsabteilung), Austausch von Sprachdaten

Mit Prof. Nobuaki Minematsu, Universität Tokio ist Kooperation bezüglich der Alterserkennung von Sprechern geplant. Seine Gruppe hat einen japanischen Korpus analysiert.

[1] Es ist uns gelungen, Zhiping Zheng, den Entwickler des Systems AnswerBus nach Saarbrücken zu holen und AnswerBus im Projekt COLLATE weiterzuentwickeln (siehe Abschnitt II.1.3).

Im Gebiet von Information Retrieval gab es Kooperationen mit Katja Markert, Universität Edinburgh (Metonymie in Information Retrieval), Ed Hovy, ISI (Marina del Rey, CA), Jaime Carbonell, CMU (Pittsburgh, PA), Mark Light, Mitre (Bedford, MA). Mit Michael Kohlhase, Carnegie Mellon University (Pittsburgh, PA), wurden Fragen der Ontologiemodellierung bearbeitet. In Bezug auf die Framenet-Annotation gab es Kooperationen mit Charles Fillmore, ICSI (Berkeley, CA). und Hans Boas, ICSI (Berkeley, CA) und University of Texas (Austin, TX), und mit Christiane Fellbaum (Princeton) in Bezug auf die Struktur des Lexikons.

In Bezug auf Korpusannotation gab es enge Zusammenarbeit mit dem Projekt TIGER und den Universitäten Potsdam und Stuttgart. In Bezug auf semantische Annotation wurden Gespräche geführt mit Joseph van Genabith, Dublin City University, und Ulrich Heid, Universität Stuttgart. Mit Adam Kilgarriff, University of Brighton gab es einen Erfahrungsaustausch zum Thema automatische Annotation großer Korpora. Im Bereich von Parsing und Integration von tiefer und flacher Verarbeitung gab es Kontakte mit Steven Abney (University of Michigan) und Stephan Oepen, YY Technologies und Stanford University).

Bei einem Besuch von Prof. Matsumoto (Nara Institute of Science and Technology, Japan) wurde das Projekt vorgestellt und Möglichkeiten der Zusammenarbeit diskutiert. Im November 2001 wurde das Projekt einer Delegation der Shanghai Jiao Tong University vorgestellt. Tianfang Yao besuchte im November 2002 die Fakultät Informatik an Shanghai Jiao Tong University, Shanghai, China zum Zwecke einer Forschungskooperation über chinesische Informationsextraktion. Dabei wurde auch eine gemeinsame Projektbewerbung für EXPO 2010 mit chinesischen Kollegen besprochen.

Es bestehen enge Kontakte mit dem Institut für Rechtsinformatik (Prof. Maximilian Herberger) an der Universität des Saarlandes, mit dem Ziel, Methoden zur Verbesserung von Information Retrieval in juristischen Texten mit Hilfe von Sprachverarbeitungsmethoden zu erforschen. Diese F&E-Aktivität wird auch in Kooperation mit der Juris GmbH (Reinhard Walker), dem führenden deutschen Anbieter für juristische Datenbanken - durchgeführt, mit dem Ziel, die Technologie auch im Bereich der juristischen Datenbanken einzusetzen.

Im Projekt Chorus des SFB 378 wird der in COLLATE entwickelte topologische Parser eingesetzt. Die Erstellung der FrameNet-Annotation findet in enger Kooperation mit dem Projekt SALSA an der Universität des Saarlandes (Computerlinguistik) statt. Andere Aufgaben, wie die Integration von GermaNet in das deutsche FrameNet werden gemeinsam von beiden Projekten durchgeführt. Es besteht auch eine Kooperation mit ICSI, wo das englische FrameNet entwickelt wurde.

Die Arbeiten im Bereich der Dialogmodellierung bauen auf den Ergebnissen der EU-Projekte TRINDI und Siridus auf. Die hier entwickelten Konzepte bilden auch die Basis für den Prototyp eines Dialogsystems, das die CLT Sprachtechnologie GmbH im Zusammenarbeit mit BMW und der TU München entwickelt. Die Ergebnisse des Projekts COLLATE fließen auch in das EU-Projekt TALK ein, das von Prof. Pinkal koordiniert wird.

Weitere Kontakte bestehen mit Martha Palmer, University of Pennsylvania im Bereich der automatischen semantischen Annotation von Korpora, PropBank und VerbNet, sowie mit Joachim Niehren, Université de Charles de Gaulle, Lille, Frankreich zu den Themen Effizienz und Ambiguität beim Parsing.

Im Bereich der Question-Answering-Systeme bestehen mehrere Kooperationen mit akademischen und industriellen Partnern. Gemeinsam mit dem Forschungszentrum Telekommunikation Wien (Ed Schofield) wurde F&E an gesprochener Ein- und Ausgabe von Question-Answering-Systemen durchgeführt. Die DFKI spin-off-Firma XtraMind hat den Einsatz unseres Question-Answering-Systems für Wissensmanagement-Produkte mit eigenen Daten gemeinsam mit COLLATE evaluiert. Monrai Technologies, Inc. ist interessiert an einer gemeinsamen Produktentwicklung von natürlichsprachlichen Question-Answering-Systemen. Die Masaryk-

Universität, Tschechische Republik ist an einer gemeinsamen Entwicklung von Question-Answering-Systemen für den Bereich medizinischer Texte interessiert. Das American National Corpus Project, USA, setzt unseren Satzsegmentierer ein; die Firma Terra Lycos hat ebenfalls Interesse an dieser Technologie bekundet. Enge wissenschaftliche Kontakte gibt es mit Prof. Sanda Harabagiu, University of Texas at Dallas im Bereich Information Retrieval und Question Answering.

Mitarbeiter des Projekts waren an nationalen und internationalen Standardisierungsaktivitäten beteiligt. Prof. Pinkal nahm auf nationaler Ebene an der konstituierenden Sitzung des DIN-Unterausschusses Sprachressourcen teil, und auf internationaler Ebene an der SIGSEM Working Group on the Representation of Multimodal Semantic Information.

II Eingehende Darstellung

II.1 Erzielte Ergebnisse

In der anwendungsorientierten Grundlagenforschung an der Universität des Saarlandes, die vom DFKI koordiniert wurde, wurden in drei Bereichen ausgewählte sprachtechnologische Verfahren und Funktionalitäten ausgearbeitet, die ein großes wirtschaftliches Potenzial haben, mit dem Ziel, sie für realistische Anwendungen nutzbar zu machen. Diese Forschung profitiert von der hervorragenden IT-Ausstattung, die durch COLLATE finanziert wurde. Die Forschung an der Universität des Saarlandes hatte drei Anwendungsschwerpunkte:

- die Extraktion von Information aus großen Mengen digitaler Texte
- natürliche Interaktion mit elektronischen Informationsdiensten
- Management großer Informationsmengen durch verbesserte Suche und Textzusammenfassung

Die wichtigsten Forschungsergebnisse des Vorhabens sind:

- SProUT, ein unifikationsbasiertes Finite-State-Toolkit für die Informationsextraktion, mit einer Benutzerschnittstelle für die Grammatikentwicklung und Evaluationswerkzeugen.
- Eine neue Methodologie für die Entwicklung multilingualer Grammatiken, angewandt auf die Erstellung von "Named Entity"-Grammatiken für acht Sprachen
- Annotation eines Korpus von Nachrichtenmeldungen mit "Named Entities", domänenspezifischen Relationen und Koreferenz-Information
- Entwicklung und Evaluation von maschinellen Lernverfahren zur Verbesserung der Informationsextraktion, angewandt auf chinesische Zeitungstexte zum Thema Sport
- Eine sprachtechnologische Architektur (framework) für Informationsmanagement auf Basis von partiell resolvierten Dependenzstrukturen (PReDS) und FrameNet-basierten semantischen Annotationen
- praktische Anwendung eines Informationszustandsmodells (Information State Update) für das Dialogmanagement
- Architektur und Implementation von M3I, einer mobilen, multimodalen und modularen Benutzerschnittstelle, einschließlich einer Medienfusionskomponente
- Entwicklung und Evaluation von Verfahren zur Bestimmung von Sprechereigenschaften (Alter und Geschlecht) aus der akustischen Eingabe
- Verbesserungen der Genauigkeit und Effizienz des web-basierten Question-Answering-Systems AnswerBus
- Anwendung der Technologien aus AnswerBus auf CNN-Nachrichten, technische Dokumentation und medizinische Dokumentsammlungen.
- Entwicklung eines effizienten Textzusammenfassers und erfolgreiche Teilnahme an der DUC2003-Evaluation.
- Methoden und Werkzeuge zum Aufbau von Webkorpora, angewandt auf den Bereich der Sprachtechnologie, ausgehend von der am DFKI erstellen Dokumentsammlung LT World.

Das Projekt wurde im Februar 2003 in Saarbrücken durch den internationalen wissenschaftlichen Beirat und den Projektträger positiv begutachtet, und die Erreichung der bis dahin vorgesehenen Meilensteine festgestellt. Eine weitere Begutachtung fand im August 2003 auf Basis eines ausführlichen englischsprachigen Projektberichts durch internationale, vom BMBF ausgewählte, Gutachter statt. Dabei wurde die erfolgreiche Durchführung des Projekts bestätigt und die Förderung des Anschlussprojekts COLLATE II befürwortet. Im folgenden werden die Ergebnisse der einzelnen Arbeitspakete dargestellt.

II.1.1 Informationsextraktion

Ergebnis der Forschung in COLLATE ist die Entwicklung unifikationsbasierten Finite-State-Formalismus (SProUT) und von multilingualen Named-Entity-Grammatiken in diesem Formalismus. Auf Basis einer erweiterten Fassung des MUC7-Standards haben wir Named-Entity-Grammatiken für Deutsch, Chinesisch, Japanisch, Französisch, Spanisch, Englisch und Tschechisch entwickelt. Die Grammatiken erkennen Personennamen, Organisationen, geographische Lokationen, Währungs-, Zeit- und Datumsausdrücke. Teilgrammatiken und Gazeteers werden soweit wie möglich für die verschiedenen Sprachen gemeinsam entwickelt. Annotierte multilinguale Korpora mit Wirtschaftsnachrichten werden für die Entwicklung und Evaluation verwendet. Das Annotationsformat für Named Entities und weitere linguistische Informationen ist eine Weiterentwicklung existierender Formate. Ein weiteres Ergebnis ist ein Evaluationswerkzeug, welches umfangreiche Statistiken und Diagnosen liefert, partielles Matching von Annotationen ermöglicht, und benutzerdefinierte Abbildungen zwischen verschiedenen Annotations- und Analyseformaten unterstützt.

II.1.1.1 Das Informationsextraktionssystem SPROUT

II.1.1.1.1 XTDL – Der Formalismus von SProUT

XTDL kombiniert zwei bewährte Ansätze: getypte Merkmalstrukturen und reguläre Ausdrücke. XTDL basiert auf TDL, einer Definitionssprache für getypte Merkmalstrukturen (Krieger und Schäfer, 1994) Das folgende TDL-Fragment, einschließlich Koreferenz und Funktionaler Applikation, ist in XTDL implementiert:

```
type-def     ::= type { avm-def | sub-def } "."
type    ::= identifier
sub-def      ::= ":<" type
avm-def      ::= ":=" avm
avm     ::= term { "&" term }*
term    ::= type | fterm | string | coref
fterm   ::= "[" [ attr-val { "," attr-val }* ] "]"
attr-val ::= attribute avm
attribute    ::= identifier
coref   ::= "#"identifier
```

In der folgenden Definition, erbt der Typ *morph* die Attribute von dem Typ *sign* und führt drei weitere morphologisch motivierte Attribute mit getypten Werten ein.

```
morph := sign & [ POS  atom, STEM atom, INFL infl ].
```

Abb. 1 zeigt ein Fragment der Typhierarchie.

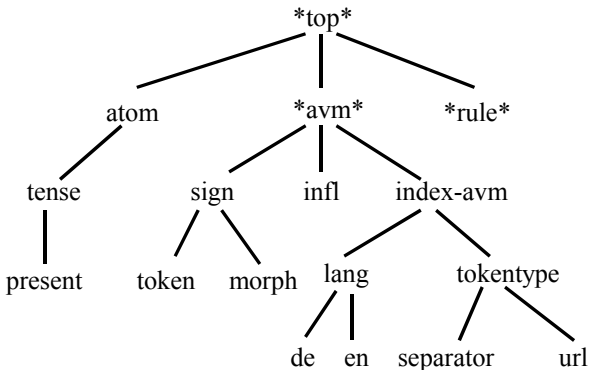

Abb. 1: Typhierarchie

Eine Regel in XTDL besteht aus einem Erkennungsmuster aus der linken Seite, geschrieben als regulärer Ausdruck, und einer Ausgabebeschreibung auf der rechten Seite. Operationen für Konkatenation, Disjunktion, Kleene-star, Kleene-plus und Optionalität stehen zur Verfügung, dargestellt durch die Operatoren |, *, +, and ?,

```
rule   ::= identifier ":>" regexp "->" {fterm}* [fun-op]"."
regexp  ::=   avm | "(" regexp ")" | regexp {regexp}+ | regexp "|" {regexp}+ | regexp { "*"
| "+" | "?" } |
           regexp "{" int [ "," int ] "}"
fun-op   ::= "where" { coref "=" fun-app }+
fun-app ::= identifier "(" term { "," term }* ")"
```

Die folgende XTDL-Grammatikregel illustriert die Syntax.

```
np :>
  (morph & [ POS  Determiner, INFL  [CASE #1, NUM #2, GEN #3 ]] )?
  (morph & [ POS  Adjective, INFL  [CASE #1, NUM #2, GEN #3 ]] )*
  (morph & [ POS  Noun & #4, INFL  [CASE #1, NUM #2, GEN #3 ]] ){1,2}
 -> phrase & [CAT #4, AGR agr & [CASE #1, NUM #2, GEN #3 ]].
```

Durch den Einsatz von TDL ergeben sich eine Reihe von Vorteilen. Getypte Merkmalstrukturen bieten eine mächtige Beschreibungssprache für linguistische Strukturen und ermöglichen Generalisierungen über reine atomare Symbole. Unifizierbarkeit als Kriterium für die Anwendbarkeit eines Zustandsübergangs kann als eine Generalisierung über die Identität von Symbolen betrachtet werden. Mit Kofererenzen in Merkmalstrukturen kann strukturelle Identität ausgedrückt werden. Koreferenzen ermöglichen dynamische Wertzuweisungen in den Übergängen von Automaten und durchbrechen die strikte Lokalität, die Ansätze mit atomaren Symbolen auszeichnet. Darüber hinaus dienen Koreferenzen als Mittel für den Transport von Information in die Ausgabebeschreibung auf der rechten Regelseite. Nicht zuletzt vereinfacht die Wahl von Merkmalstrukturen als Objekte die Komposition von NLP-Modulen, da alle Ein- und Ausgaben denselben abstrakten Datentyp verwenden.

II.1.1.1.2 Systembeschreibung

Der Kern des Systems SProUT besteht aus den folgenden Komponenten:

- eine Menge von Werkzeugen zur Konstruktion, Kombination und Optimierung verschiedener Typen von endlichen (finite-state) Automaten,
- ein flexibler XML-basierter Compiler, der reguläre Ausdrücke in die entsprechende komprimierte Finite-State-Repräsentation überführt (Piskorski et al., 2002),
- das Java-Package JTFS, das Standardoperationen für die Konstruktion und Manipulation getypter Merkmalstrukturen bereitstellt, und
- der Grammatikinterpreter XTDL.

SProUT enthält drei linguistische Verarbeitungskomponenten: einen Tokenizer, Gazetteer, und eine morphologische Analysekomponente für jede Sprache. Die Architektur von SProUT wird in Abb. 2 dargestellt.

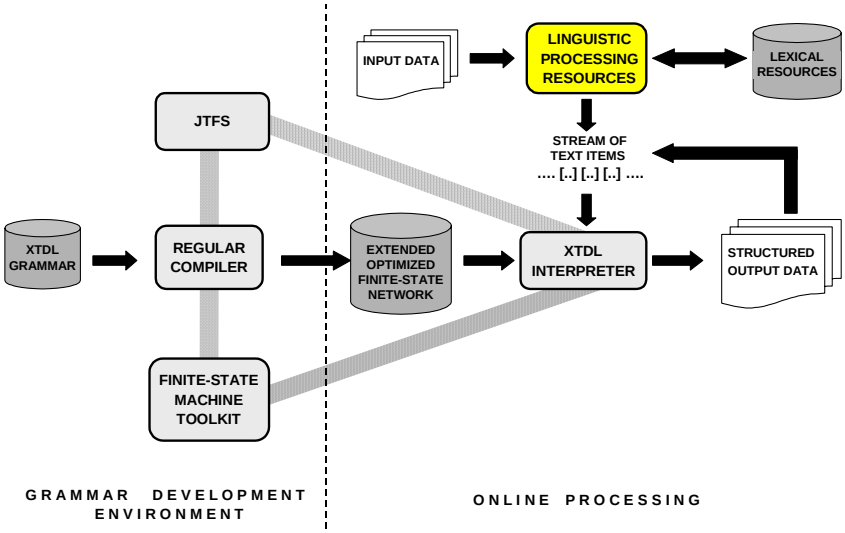

Abb. 2: Architektur von SProUT

SProUT hat eine benutzerfreundliche Entwicklungsumgebung für Grammatikschreiber. Der Entwicklungsprozess beginnt mit einer Projektdefinition. Ein Projekt besteht aus einer Grammatikdefinition und einer Systemkonfiguration, sowie einer extern definierten Typenhierarchie.

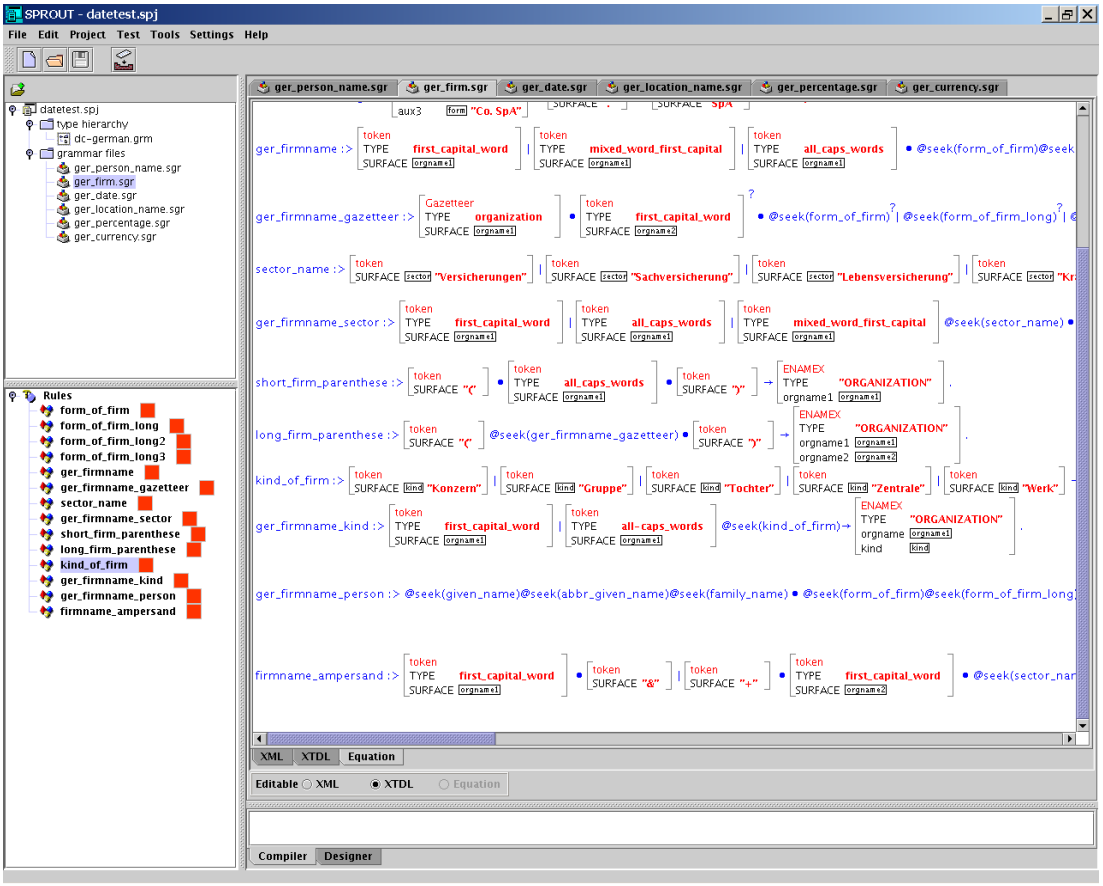

Abb. 3: Grammatikentwicklungsumgebung von SProUT

Die Systemkonfiguration erlaubt die Spezifikation von einzelnen Verarbeitungskomponenten und den dazugehörigen Ressourcen. Die Grammatik kann mit Eingabetexten getestet werden, wobei die verwendeten und instantiierten Regeln und der abgedeckte Text hervorgehoben werden. Abb. 4 zeigt das Testen von japanischen Named-Entity-Grammatiken. Das System ist in Java und C++ implementiert und läuft unter den Betriebssystemen Windows und Linux.

SProUT wurde auch in den EU-geförderten Verbundprojekten AIRFORCE und MEMPHIS eingesetzt, die am DFKI durchgeführt wurden. In AIRFORCE wurde SProUT zur Extraktion von Information aus Reisewarnungen in mehreren Sprachen verwendet. Einzelne Komponenten des SProUT-Systems wurden gemeinsam mit dem Projekt MEMPHIS weiterentwickelt.

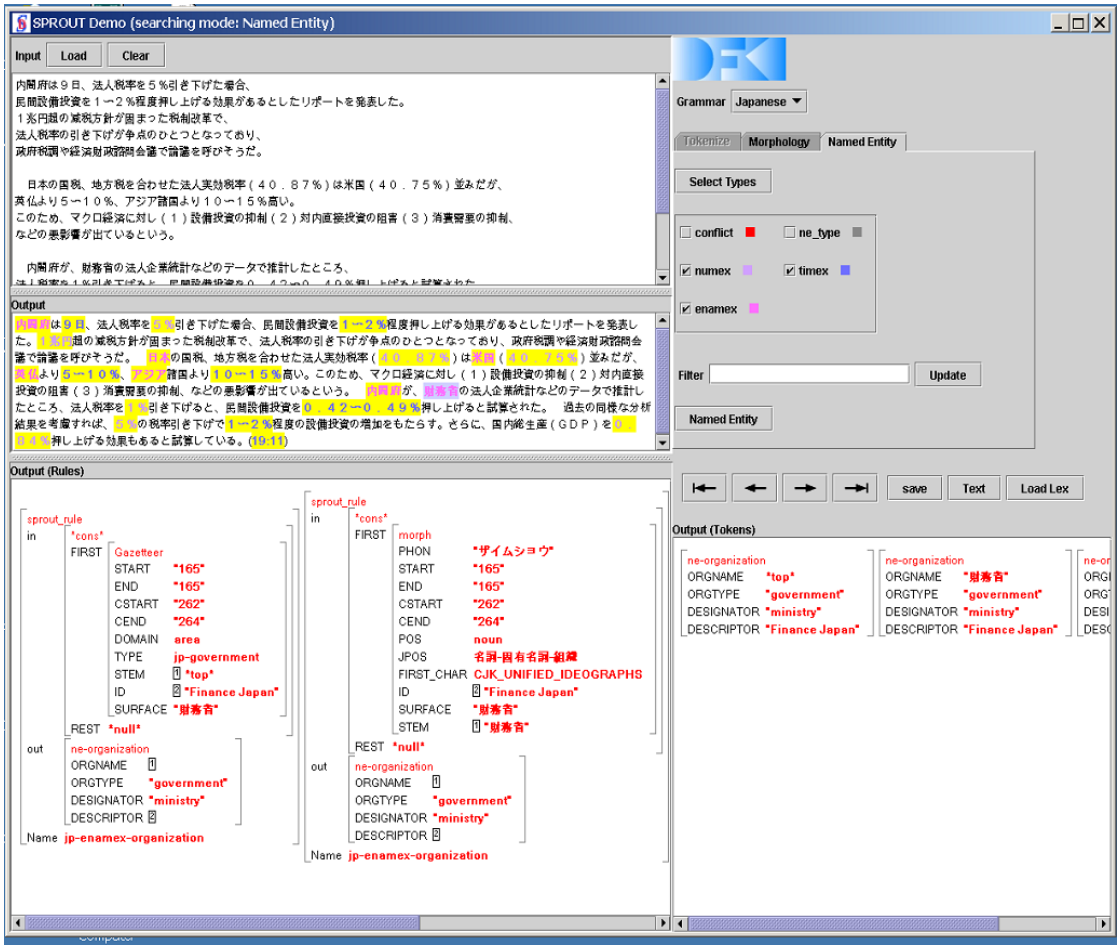

Abb. 4: Der multilinguale Named-Entity-Erkenner

II.1.1.2 Multilinguale Grammatikentwicklung

Ein wichtiges Ziel bei der Entwicklung von multilingualen Named-Entity-Grammatiken ist die maximale Wiederverwendung von Ressourcen für verschieden Sprachen. Tokenklassen, Ausgabestrukturen und Fragmente der Grammatik werden für verschiedene Sprachen gemeinsam entwickelt und genutzt. Dadurch wird auch die Wartbarkeit und Konsistenz der linguistischen Ressourcen erhöht. Die Named-Entity-Grammatiken für alle Sprachen basieren auf einer gemeinsamen Typenhierarchie, die den MUC7-Standard erweitert.

II.1.1.3 Evaluationswerkzeuge

Zur Evaluation der Grammatiken mittels annotierter Korpora wurde die Anwendung jTaCo entwickelt. jTaCo entfernt alle Annotationen aus dem Korpus und übergibt die unannotierten Texte der Grammatik. Es vergleicht dann die Ausgabe der Grammatik mit den ursprünglichen annotierten Grammatiken und liefert detaillierte Statistiken, quantitative Auswertungen und diagnostische Ausgaben, die die Unterschiede zwischen der Ausgabe der Grammatik und der ursprünglichen Korpusannotation hervorheben. Die Architektur von jTaCo ist in der untenstehenden Abbildung dargestellt. jTaCo ist in einem grafischen Benutzerinterface konfigurierbar und kann mit verschiedenen Problemen bei der Evaluation von Grammatiken im Vergleich zu annotierten Korpora umgehen:

- Verwendung verschiedener Bezeichnungen für Klassen von Named Entities, oder unterschiedliche Granularität der Grammatik und der Korpusannotation (z.B. O*rganisation* und die Unterklassen *Firma, Universität, Regierung* usw.)
- Die Spanne einer Named Entity kann unterschiedlich sein; z.B. kann ein Personenname mit oder ohne Funktion und Titel annotiert bzw. extrahiert werden ("Chairman and CEO Bill Gates" vs. "Bill Gates")
- Die Korpusannotation kann textorientiert sein (z.B. mit XML tags im Text) während die Ausgabe der Grammatik eine semantische Struktur sein kann.

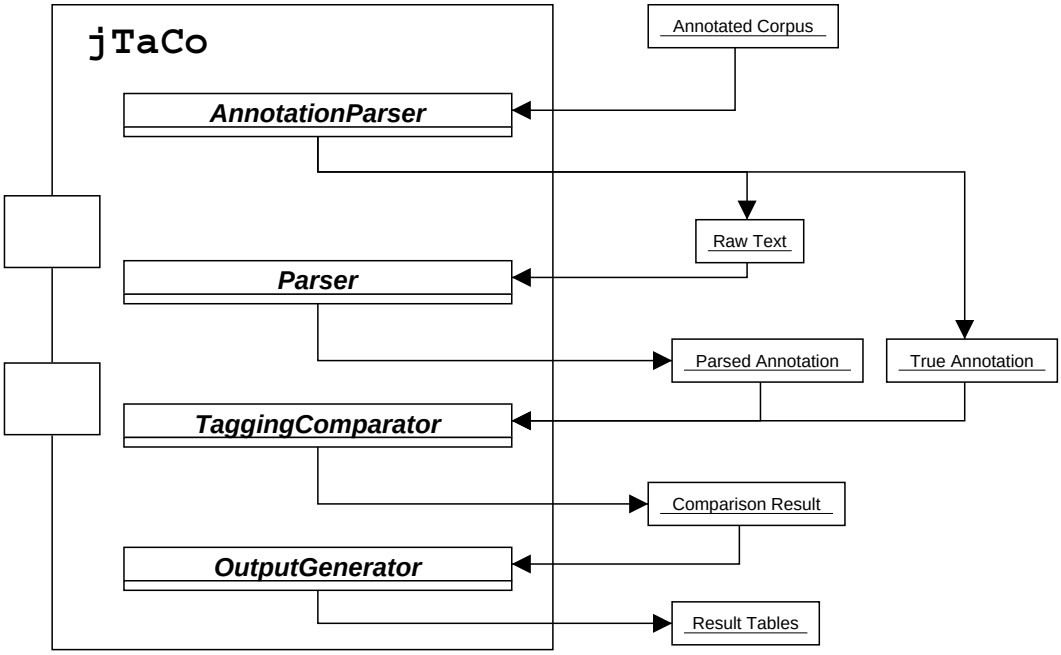

Abb. 5: Architektur von jTaCo

II.1.1.4 Evaluationsergebnisse

Die Abdeckung der deutschen Named-Entity-Grammatiken wurde anhand von annotierten Wirtschaftsmeldungen (mehr als 14.000 Wörter) aus drei deutschen Zeitungen evaluiert. Precision und Recall für jede Art von Named Entity und für jedes Zeitungskorpus sind in Tabelle 1 aufgeführt.

	Frankfurter Rundschau		Wirtschaftswoche		Süddeutsche Zeitung	
NE	**precision**	**recall**	**precision**	**recall**	**precision**	**recall**
Organization	85,71	68,57	95,65	72,13	94,07	79,12
Location	86,95	69,44	96,77	83,33	98,47	80,16
Currency	100,00	77,77	100,00	88,57	100,00	88,88
Time date	89,47	89,47	88,88	76,19	99,00	86,20
Person	83,33	71,42	75,00	60,00	59,37	82,60
Percentage	100,00	85,71	100,00	93,33	100,00	86,36
Total	89,47	70,83	94,84	80,36	95,66	82,25

Tabelle 1: Evaluationsergebnisse für Named-Entity-Erkennung

II.1.1.5 Maschinelle Lernverfahren für Informationsextraktion

Anhand des Problems der domänenspezifischen Informationsextraktion für chinesische Sportnachrichten wurden maschinelle Lernverfahren entwickelt und erprobt. Es wurden verschiedene Strategien für Informationsextraktion aus chinesischen Texten untersucht und die Nützlichkeit von maschinellen Lernverfahren für die Korrektur von Fehlern bei der Wortsegmentierung und bei der Wortartenzuweisung (part-of-speech tagging) gezeigt.

Hauptergebnis dieser Arbeiten ist die Architektur und Implementation des Systems CHINERS (Chinese Named Entity Recognition System) in Java als Teil des chinesischen Informations-extraktionssystems CIEPS. CHINERS basiert auf maschinellen Lernverfahren und einem flachen Parser. Sechs Arten von Named Entities können erkant werden: *personal name* (*PN*), *date/time* (*DT*), *location name* (*LN*), *team name* (*TN*), *competition title* (*CT*), *personal identity* (*PI*).

Ein neues Verfahren des Maschinenlernens, das „Positive and Negative Case-Based Learning“ (PNCBL) heißt, wurde vorgeschlagen. Das Verfahren gehört zu „supervised statistical learning“ und ist eine Variante von „memory-based learning“. Aufgrund dieses Verfahrens wurden zwei entsprechende Algorithmen entworfen und implementiert: ein maschinelles Lernverfahren und ein Erkennungsalgorithmus für Named-Entity-Relationen. Durch das maschinelle Lernverfahren wurden Bibliotheken von Relationsmustern und Nichtrelationsmustern aufgebaut.

Die entwickelte Sport-Ontologie hat 223 Objektsknoten, 346 Bewegungsknoten und 24 Eigenschaftsknoten mit entsprechender Information.

In der letzten Projektphase wurde das System CHINERRS (Chinese Named Entity and Relation Recognition System) entwickelt, das das vorher verwirklichte System CHINERS und eine neue Komponente zur Erkennung von Named-Entity-Relationen einschließt.

Es wurden mehrere Evaluationen des Systems durchgeführt. Dazu wurden Texte aus der Sportdomäne gesammelt und auf Basis einer Ontologie mit 12 verschiedenen Relationen zwischen Named Entities in einem XML-Format annotiert. Es wurde ein durchschnittlicher Recall von 83% und durchschnittliche Präzision von 85% erreicht.

II.1.1.5.1 Evaluation

Zur Evaluation des Systems wurden verschiedene Experimente durchgeführt. Das erste evaluiert nur die Leistungsfähigkeit der Fehlerkorrekturkomponente. Das zweite vergleicht die Erkennung von Named Entities mit und ohne Fehlerkorrektur. Das dritte untersucht die Erkennung von Team-Namen (TN) und Wettbewerbs-Namen (CT) mit und ohne Schlüsselwörter. Die Trainingsdaten bestehen aus 94 Texten mit 3473 Sätzen (37077 Zeichen) aus der Zeitung *Jiefang Daily* (2001). Die Evaluationsergebnisse sind in den folgenden Tabellen angegeben.

	PN	**DT**	**LN**	**TN**	**CT**	**PI**	**Total**
ohne Fehlerkorrektur	37%	69%	49%	39%	66%	85%	58%
mit Fehlerkorrektur	74%	93%	71%	82%	87%	93%	83%

Tabelle 2: Recall-Ergebnisse

	PN	**DT**	**LN**	**TN**	**CT**	**PI**	**Total**
ohne Fehlerkorrektur	31%	93%	52%	65%	71%	79%	65%
mit Fehlerkorrektur	67%	97%	80%	91%	88%	89%	85%

Tabelle 3: Precision-Ergebnisse

	Total Number	Total Recognized Number / (Total Error Number)	**Average Recall**	Average **Precision**
TN without keyword	**65**	**56 / (19)**	**86.15**	**66.07**
CT without keyword	**45**	**44 / (1)**	**97.78**	**97.73**

Tabelle 4: Erkennungsleistung für TN und CT ohne Schlüsselwort

Die wesentlichen Ressourcen, die für das Maschinenlernen und die Erkennung von Named-Entity-Relationen nutzbar sind, sind die Bibliotheken von Relationsmustern und Nichtrelations-mustern. Während des Maschinenlernens werden beide Bibliotheken abhängig von den kommentierten Texte aus Jie Fang Daily in 2001 und der Sport-Ontologie aufgebaut. Die zwei Bibliotheken beinhalten 142 (534 Relationen) und 98 (572 Nichtrelationen) Satzgruppen.

Die Resultate des ersten Experiments werden in Tabelle 5, Tabelle 6 und Tabelle 7 gezeigt. Die erste und zweite Tabelle bezeichnen die durchschnittliche Recall und Precision für 14 Relationen. Die dritte Tabelle stellt die gesamte durchschnittliche Recall, Precision und F-Measure für 14 Relationen dar. Das Experiment wurde nur durch „Positive Case-Based Learning and Recognition“ durchgeführt.

LOC_CPC	**PS_ID**	**PS_TM**	**TM_CP**	**TM_CPC**	**ID_TM**	**CP_LOC**
100	100	100	100	100	90.91	88.89
CP_TI	**WT_LT**	**HT_VT**	**PS_CP**	**PS_CPC**	**CP_DA**	**DT_DT**
83.33	80	71.43	60	33.33	0	0

Tabelle 5: Durchschnittlicher Recall für 14 Relationen (nur durch “Positive Case-Based Learning and Recognition”)

LOC_CPC	**PS_ID**	**PS_TM**	**TM_CP**	**TM_CPC**	**ID_TM**	**CP_LOC**
91.67	84.62	72.73	87.50	42.50	66.67	69.70
CP_TI	**WT_LT**	**HT_VT**	**PS_CP**	**PS_CPC**	**CP_DA**	**DT_DT**
71.43	30.77	38.46	75	66.67	0	0

Tabelle 6: Durchschnittliche Precision für 14 Relationen (nur durch “Positive Case-Based Learning and Recognition”)

Total Average Recall	71.99
Total Average Precision	56.98
Total Average F-measure	63.61

Tabelle 7: Gesamtdurchschnittliche Recall, Precision und F-Measure für 14 Relationen (nur durch “Positive Case-Based Learning and Recognition”)

Im zweiten Experiment wurde „Positive and Negative Case-Based Learning and Recognition“ evaluiert. Die Tabelle 8 und Tabelle 9 illustrieren die durchschnittliche Recall und Precision für 14 Relationen. Tabelle 10 beschreibt die gesamte durchschnittliche Recall, Precision und F-Measure für 14 Relationen.

CP_DA	CP_TI	LOC_CPC	PS_CPC	TM_CP	ID_TM	CP_LOC
100	100	100	100	100	90.91	88.89
PS_TM	**PS_ID**	**DT_DT**	**PS_CP**	**WT_LT**	**HT_VT**	**TM_CPC**
80	72.22	66.67	60	60	42.86	37.50

Tabelle 8: Durchschnittlicher Recall für Relationen (durch "Positive and Negative Case-Based Learning and Recognition")

CP_DA	CP_TI	LOC_CPC	PS_CPC	TM_CP	ID_TM	CP_LOC
50	75	91.67	68.75	87.50	68.19	66.67
PS_TM	**PS_ID**	**DT_DT**	**PS_CP**	**WT_LT**	**HT_VT**	**TM_CPC**
65	81.67	66.67	75	37.50	30	31.25

Tabelle 9: Durchschnittliche Precision für 14 Relationen (durch "Positive and Negative Case-Based Learning and Recognition")

Total Average Recall	78.50
Total Average Precision	63.92
Total Average F-measure	70.46

Tabelle 10: Gesamtdurchschnittliche Recall, Precision und F-Measure für 14 Relationen (durch "Positive and Negative Case-Based Learning and Recognition")

Die experimentellen Resultate zeigen, dass das Verfahren bessere Leistungen für die Erkennung der Named Entities erreicht. Außerdem zeigt ein Vergleich der beiden experimentellen Resultate begreifen, dass das gesamte durchschnittliche F-Measure durch Verwendung von beiden Fälle (positive and negative cases).von 63.61% auf 70.46% erhöht wird.

II.1.1.6 Bibliographie

[Asahara, 2000] M. Asahara and Y. Matsumoto. *Extended Models and Tools for High-Performance Part-of-Speech Tagger*, In Proceedings of the 18th COLING, pages 21-27, 2000.

[Becker et. al, 2002] M. Becker, W. Drożdżyński, H.U. Krieger, J. Piskorski, U. Schäfer, F. Xu. *SProUT - Shallow Processing with Typed Feature Structures and Unification*. In Proceedings of ICON 2002 - International Conference on NLP, Mumbai, India, December, 2002.

[Chinchor 1997] Nancy Chinchor. *MUC7 Named Entity Task Definition.* Technical Report, NIST

[Callmeier et al. 2002] U. Callmeier, G. Erbach, I. Gogelgans, S. Hansen, K. Kunz and D. Ziegler-Eisele. *COLLATE Annotationsschema.* Technical Report, Saarland University, 2002.

[Krieger and Schäfer, 1994] H.-U. Krieger, U. Schäfer. *TDL – A Type Description Language for Constraint-Based Grammars*. In Proceedings of COLING, pages 893-899, 1994.

[Liu, 2001] K. Liu. *Research of automatic Chinese word segmentation*. In International Workshop on Innovative Language Technology and Chinese Information Processing (ILT&CIP-2001), 2001.

[Petitpierre and Russell, 1995] D. Petitpierre and G. Russell. *MMORPH-The Multext Morphology Program*, 1995. Multext deliverable report 2.3.1. ISSCO, University of Geneva.

[Piskorski et. al, 2002] J. Piskorski, W. Drożdżyński, F. Xu, O. Scherf. *A Flexible XML-based Regular Compiler for Creation and Converting Linguistic Resources*. In Proceedings of LREC 2002.

[Piskorski, 2002] J. Piskorski. *DFKI Finite-State Machine Toolkit*. Research Report RR-02-04, DFKI GmbH - German Research Center for Artificial Intelligence, Saarbrücken, Germany, 2002.

[Piskorski and Neumann, 2000] J. Piskorski, G. Neumann. *An Intelligent Text Extraction and Navigation System.* Proceedings of RIAO - Content-Based Multimedia Information Access, Paris, 2000

II.1.2 NLP-Framework für Informationsmanagement

Sprachtechnologie kann einen substantiellen Beitrag leisten zu Systemen, die den präzisen und benutzerorientierten Zugriff auf große Informationsmengen in elektronischer Form unterstützen. Wir haben zwei Dimensionen dieses Problemraums untersucht, nämlich die Verbesserung der Suchverfahren für Question Answering (QA) und Informationsextraktion (IE) durch linguistische Verfahren und Unterstützung der Benutzerinteraktion mit einem Informationssystem mittels Dialog.

Linguistische Information ist ein Schlüsselfaktor bei der Entwicklung von QA- und IE-Systemen, die leistungsfähiger und vielseitiger sind beim zielgerichteten Auffinden der benötigten Informationen, wie die Erfahrung mit Systemen für die englische Sprache gezeigt hat. Voraussetzung für diese Entwicklungen ist eine robuste, anpassbare und domänen-unabhängige Verarbeitungskomponente für natürliche Sprache, die zum Projektbeginn für die deutsche Sprache noch nicht verfügbar war. Weitere Verbesserungen werden durch Hinzufügen einer Ebene der lexikalisch-semantischen Repräsentation erwartet. In COLLATE haben wir an folgenden Aufgaben gearbeitet:

- Implementation einer robusten Komponente für die Verarbeitung großer Menge deutscher Texte bis zur syntakto-semantischen Ebene, die als Grundlage für die Entwicklung von praktischen Anwendungen dienen kann.
- einem robusten Parser und seiner Anwendung für Informationsextraktion und Question Answering, wobei die Ausgabe des Parsers für die automatische Generierung einer Ebene der semantischen Repräsentation verwendet wurde.

II.1.2.1 Parsing deutscher Texte in syntakto-semantische Strukturen

Für die Zielapplikationen Question Answering und Informationsextraktion wird eine linguistische Repräsentation benötigt, die mächtig genug ist, um maschinelles Schließen jenseits der Ebene der Oberflächenrealisierung zu ermöglichen. Andererseits muss die Repräsentation einfach genug sein, dass sie zuverlässig für große Mengen von Text aus verschiedenen Domänen erzeugt werden kann. Um dieses Ziel zu erreichen, haben wir die Repräsentationssprache PReDS (Partially Resolved Dependency Structure) entworfen und eine Prozesskette von NLP-Werkzeugen implementiert, die diese PReDS-Strukturen aus Texten extrahieren.

II.1.2.1.1 Architektur

Die Architektur der Verarbeitungskette folgt dem Prinzip des *Easy-first Parsing.* Dieses Konzept wurde von Abney (1996) eingeführt und beschreibt eine sequentielle NLP-Architektur, in der einfache Komponenten aneinander gekoppelt werden, wobei jede Komponente nur Aktionen ausführt, die zuverlässige Ergebnisse liefern, und schwierigere Fälle späteren Komponenten überlässt, die mehr Informationen zu deren Lösung zur Verfügung haben. Damit ergibt sich ein sehr robustes Parsingsystem, das Ausgaben erzeugt, die nicht immer vollständig sind, dafür aber zuverlässig und gut anwendbar.

Unser Parsingsystem besteht aus den folgenden vier Komponenten, die im Folgenden genauer beschrieben werden:

- Tokenisierung und Morphologie
- Topologischer Parser
- Phrasaler Chunker
- PReDS-Generator

Alle Komponenten kommunizieren durch XML-basierte Austauschformate.

II.1.2.1.2 Topologischer Parser

Parsing von natürlichen Sprachen ist eine sehr komplexe Aufgabe. Durch eine Divide-and-Conquer-Stategie [Peh and Ting 1996] wird das Gesamtproblem aufgeteilt in kleinere, handhabbare Teile. Die Topologie des deutschen Satzes ist ein perfekter Ausgangspunkt für eine solche Strategie, da sie konsistent und konzeptuell einfach ist, jedoch eine effektive und nützliche Methode zum Strukturieren von Sätzen.

Topologische Grammatik

Da die Menge der topologischen Regeln sehr eingeschränkt ist (jeder Satz wird in fünf Teile aufgeteilt, nämlich *Vorfeld, linke Klammer, Mittelfeld, rechte Klammer, Nachfeld;* siehe. Abb. 6) und sie fast ohne Ausnahmen gelten, verwenden wir eine kompakte hand-codierte kontextfreie Grammatik für den topologischen Parser. Diese Grammatik wird von einem Earley-artigen Parser benutzt, der in Perl implementiert wurde.

Die Grammatik benutzt eine kleine Menge von verlässlichen Indikatoren (Verben, Subjunktionen, Konjunktionen und Verbpräfixe), um die topologische Struktur zu erkennen (Braun 1999). Aktuell besteht die Grammatik aus 273 Regeln, die eine breite Spanne von Konstruktionen abdecken, einschließlich verschiedener Typen von Verbkomplexen, Koordinationskonstruktionen auf verschiedenen syntaktischen Ebenen, kaskadierten Nebensätzen usw.

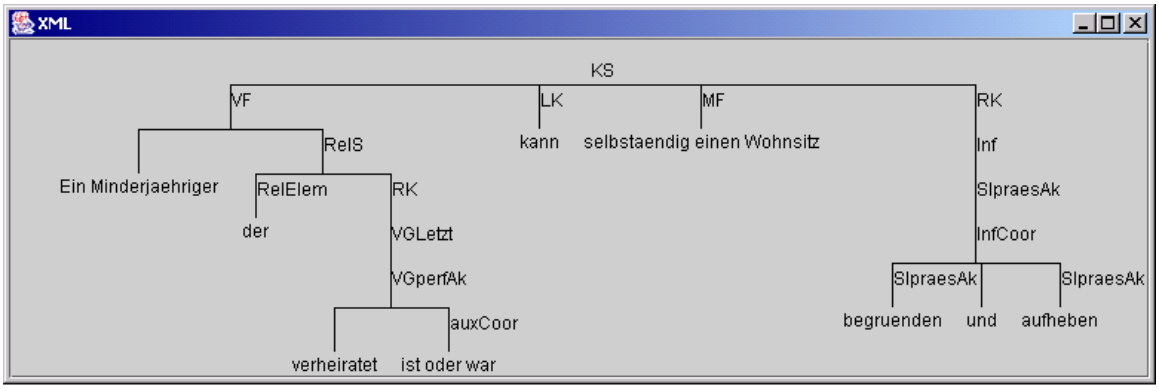

Abb. 6: Topologische Analyse

Evaluation

Um die Leistungsfähigkeit des topologischen Parsers für die Zielanwendung zu beurteilen, haben wir ihn mit einem ungesehenen Korpus aus 104 Seiten aus dem Bürgerlichen Gesetzbuch und 353 Sätzen aus Wirtschaftsmeldungen der Süddeutschen Zeitung evaluiert. Dieser Korpus wurde in 100 Sekunden geparst - der Parser verarbeitet also etwa 4 bis 5 Sätze pro Sekunde. Für die Evaluation sind wir davon ausgegangen, dass jeder Satz genau eine richtige topologische Struktur hat. Die Ergebnisse werden durch zwei Metriken ausgewertet:

- **Recall**: Verhältnis der Anzahl der korrekten Analysen des Parsers zu der gesamten Anzahl von Sätzen.
- **Precision**: Verhältnis der korrekten Analysen des Parsers zu der gesamten Anzahl von Analysen.

	Sentences				Trees			Recall	Precision
	Total	**Falsch**	**Amb**	**K.A.**	**Total**	**Correct**	**Wrong**		
SZ	353	18	23	18	360	315	45	89.24%	87.50%
BGB	104	6	7	14	100	84	16	80.77%	84.00%
Total	457	24	30	32	460	399	61	**87.31%**	**86.74%**

Falsch: eindeutige, aber falsche Analyse Amb: mehr als eine Analyse K.A.: Keine Analyse

Diese Ergebnisse illustrieren einige Punkte: Zum einen, dass juristische Texte aufgrund ihrer komplexen Struktur schwerer zu analysieren sind als Zeitungssprache. Dies zeigt sich anhand der Evaluationsergebnisse für den SZ-Korpus und den BGB-Korpus. Aber auch für juristische Texte erreicht der Parser akzeptable Ergebnisse.

Zum zweiten zeigt sich, dass strukturelle Ambiguität kein großes Problem ist, da die Anzahl der Sätze mit mehr als einer topologischen Analyse klein genug ist, dass man sie für QA- und IE-Anwendungen ignorieren kann.

Die Ergebnisse bestätigen uns in der Entscheidung für eine regelbasierte, manuell erstellte Grammatik. Nach Analyse der Fehlerquellen gibt es noch Möglichkeiten zur Verbesserung, womit die Evaluationswerte wohl in die Region von 90% gesteigert werden können. Wir haben gezeigt, dass der topologische Parser eine sehr zuverlässige Komponente der linguistischen Verarbeitungskette ist.

Auf Grundlage der topologischen Strukturen ist es sehr einfach, den Text der verschiedenen topologischen Felder zu extrahieren und an die nächste Stufe in der Verarbeitungskette weiterzugeben, die aus einem Named-Entity-Erkenner und einem NP/PP-Chunker besteht.

II.1.2.1.2.1 Named Entities

Named Entities sind extra-linguistische Spezialausdrücke, d.h. Phrasen, die nicht nach allgemeinen linguistischen Prinzipien konstruiert sind, sondern idiosynkratische Strukturen aufweisen, wie z.B. *Hellas Reisen GmbH*, *ca. 5 Mrd. Dollar* oder *§ 5 Abs. 2 RettAssG*. Diese Form von Ausdrücken ist in allen Fachtexten (seien es Wirtschaftsmeldungen, Sportberichte oder richterliche Urteile) allgegenwärtig. Wir haben ein auf regulären Ausdrücken basierendes Modul zur NE-Erkennung entwickelt und in den Verarbeitungsablauf des Parsing-Systems integriert. Es wird nach der topologischen und vor der Phrasen-Analyse angesprochen. Named-Entity-Erkennung und NP/PP-Chunking werden damit zwar in zwei getrennten Verarbeitungsschritten durchgeführt. Da NEs und rein linguistisches Parsen aber eng miteinander interagieren, haben wir den NP/PP-Chunker so erweitert, dass die Resultate der Named-Entity-Erkennung in das NP/PP-Chunking eingebunden werden. Das ermöglicht zum einen die korrekte Erkennung von NPs/PPs, die Named Entities als Konstituenten enthalten (insb. wichtig bei Koordination und komplexeren NPs). Zum anderen können damit die NE-Grammtiken völlig auf den idiosynkratischen Teil der NEs beschränkt werden, während linguistische Elemente wie z.B. Artikel oder Adjektive, die ebenfalls Teil der gesamten NE sein können, vom NP/PP-Chunker abgedeckt werden. Dadurch konnte die Erkennungsleistung beider Teilkomponenten noch einmal deutlich gesteigert werden.

Das Modul zur NE-Erkennung arbeitet zum großen Teil mit Grammatiken, die wir von der Collate IE-Gruppe übernommen und für unseren Ansatz operationalisiert haben. Lediglich

Grammatiken zur Erkennung von Gesetzeszitaten, Datumsangaben und Personennamen mussten neu entwickelt werden. Dabei kommen auch die umfangreichen, ebenfalls von der IE-Gruppe erstellten NE-Listen in Form eines Gazetteers zum Einsatz.

Als Alternative zur geschilderten regelbasierten NE-Erkennung wurde ein Verfahren untersucht, das einen Lernalgorithmus und einen durch Internet-Daten erstellten Gazetteer benutzt. Mit Hilfe einer kleinen Menge von Startelementen werden dabei auf sehr großen Textmengen aus dem WWW auf robuste und domänenunabhängige Weise umfangreiche Listen von Named Entities erzeugt. Dies eröffnet die Möglichkeit, den von den Regeln genutzten Gazetteer automatisch zu erweitern.

II.1.2.1.2.2 NPs und PPs

Der auf endlichen Automaten basierende NP/PP-Chunker nimmt die morphologische Analyse der Zwei-Ebenen-Morphologie von Lingsoft als Eingabe. NPs und PPs werden mit einer handcodierten regulären Grammatik verarbeitet. Die Grammatik wurde erweitert, um selbsteinbettende NPs und PPs zu erkennen, die im Deutschen häufig vorkommen. Da die grammatischen Regeln für deutsche NPs relativ kompakt und regelmäßig sind, verwenden wir eine handcodierte Grammatik, womit auch morphologische Merkmale und Kongruenz berücksichtigt werden können, die Schwierigkeiten für statistische Ansätze bereiten. Evaluationsergebnisse zeigen, dass der NP-Chunker eine Genauigkeit von etwa 90 % erreicht. Der Chunker wird in [Fliedner 2002] ausführlich beschrieben.

Die Ausgabe des Chunkers ist eine Liste von Phrasen, die wieder in die topologische Struktur eingefügt werden, womit Extended Topological Structures (XTS) erzeugt werden. Diese dienen als Eingabe für die nächste Komponente, den PReDS-Generator.

II.1.2.1.3 PReDS-Generator

PReDS (*Partially Resolved Dependency Structures*) dienen als Schnittstelle zwischen der robusten, domänenunabhängigen linguistischen Analyse und der feinkörnigeren Verarbeitung der nachgelagerten Module. PReDS ist eine flache syntakto-semantische Repräsentation, die linguistische Vollständigkeit und Präzision gegen Zuverlässigkeit und Eindeutigkeit eintauscht, um linguistische Anwendungen zu unterstützen.

Konzeptuell sind PReDS Prädikat-Argument-Strukturen, die nahe an der syntaktischen Oberflächenstruktur sind mit semantisch motivierten Modifikationen, z.B. wird bei Passiv-Konstruktionen das Tiefensubjekt repräsentiert. Wenn weiteres linguistisches oder kontextuelles Wissen benötigt wird, verwendet der PReDS-Generator Default-Labels. Das typische Beispiel für diese Strategie sind PPs, denn die Berechnung der korrekten PP-Anbindung übersteigt die Möglichkeiten von XTS; daher werden PPs immer so niedrig wie möglich angebunden, solange es keine zuverlässige Evidenz gibt, dass die PP zu einer spezifischen NP bzw. Verb gehört.

Ein Arbeitsschwerpunkt in der letzten Projektphase war die Erweiterung der Abdeckung des PreDS-Parsers, um ihn von einem Prototypen zu einem für realistische Texte einsetzbaren Tool weiterzuentwickeln. Mittlerweile kann der Parser mit einer großen Menge von syntaktischen Konstruktionen korrekt umgehen, wie z.B. verschiedene Formen von Koordinationen auf Haupt- und Nebensatzebene, asyndetische Konditionale, Aktiv- und Passiv-Variationen, Erkennen und Einfügen nicht realisierter Argumente in Kontrollverbkonstruktionen usw.

```
mitteilen[+sg, +past, +want]
   --DSub-->Carrier[+masc, +sg, +def]
               --Mod-->belgisch
   --ClArg-->verhandeln[+pres, +pl]
                --DSub-->und
                            --Coor-->Airline[+def, +fem, +sg]
                                        --Mod-->belgisch
                            --Coor-->Luxair[+Nameexpr, +sg ... ]
                --PPModDef-->ueber
                             --Arg-->Zusammenschluss[+sg ...]
```

Abb. 7: PReDS für einen Beispielsatz: *Die belgische Airline und Luxair wollen über einen Zusammenschluss verhandeln, teilte der belgische Carrier mit.*

II.1.2.1.4 Anwendungsszenario: Textmining juristischer Texte

Um den Nutzen des vorgeschlagenen NLP-Frameworks für QA und IE zu demonstrieren, wenden wir es für IE in der juristischen Domäne an. Die juristische Domäne erscheint als ein passender Testfall für NLP-basierte Informationsmanagement-Strategien, denn sie bietet

- große Mengen elektronisch verfügbarer Texte hoher Qualität,
- eine geschlossene, aber hinreichend komplexe Domäne, für die ausgefeilte QA- und IE-Strategien angewandt werden können,
- einen realistischen Anwendungsbedarf.

Das Extrahieren von Informationen aus freien Texten ist eine typische Aufgabe der Informationsextraktion (IE). Im Gegensatz zu konventionellen IE-Techniken, die vordefinierte Templates füllen, haben wir bei der Extraktion von Definitionen kein A-priori-Wissen über Domänen und das verwendete Vokabular. Die Aufgabe der Extraktion von Definitionen übersteigt die Fähigkeiten existierender IE-Technologien. Die flachen Verfahren, die heute in vielen QA- und IE-Systemen verwendet werden, reichen nicht aus, um Informationen zu liefern, die präzise und umfangreich genug für die intendierten Anwendungen sind. Statt dessen verwenden wir Extended Topological Structures (XTS) und definieren lexikalisch-syntaktische und kontextuelle Muster zur Extraktion von Definitionen.

Die Verfügbarkeit zentraler juristischer Begriffe und ihrer Definitionen ist von großem Nutzen für ein QA-System in der juristischen Domäne. Erstens kann für einige Fragen zielführende und präzise Information geliefert werden, zweitens ermöglichen Term-Hierarchien die Verwendung einfacher Inferenzmechanismen für Question-Answering, und drittens eröffnet sich die Möglichkeit der automatischen Konstruktion von domänenspezifischen Ontologien, wenn die Extraktionsstrategien für verschiedene Bereiche der juristischen Domäne sich als zuverlässig und konsistent herausstellen.

II.1.2.2 Lexikalische Semantik im Informationsmanagement

Wir untersuchen, wie linguistische Information angewandt werden kann, um die Leistungsfähigkeit und Funktionalität von Informationssystemen zu verbessern. Unser Schwerpunkt liegt auf flachen semantischen Repräsentationen als zusätzlicher Beschreibungsebene. Als Formalismus haben wir FrameNet-Strukturen gewählt.

II.1.2.2.1 Semantische Annotation als zusätzliche Informationsebene

Die Verwendung einer semantischen Repräsentation hilft uns, weiter von der Obenflächenstruktur der Texte zu abstrahieren. PReDS normalisieren schon einige der oberflächenorientierten Phänomene wie Aktiv und Passiv. Mit einer flachen semantischen Repräsentation normalisieren wir systematische Relationen zwischen verschiedenen Oberflächenrealisierungen, insbesondere

- (Nahe) Synonymie, z.B. *kaufen* vs. *erwerben*
- (Nahe) Hypo/Hyperonymie, z. B. *bestellen* vs. *anfordern*
- Nominalisierungen, z. B. *A verkauft B* vs. *As Verkauf von B*
- Inverse Relationen, z. B. *A kauft B von C* vs. *C verkauft B an A*
- Syntaktische Variation von Präpositionen, z. B. *Auftrag für 10 Flugzeuge* vs. *Auftrag über 10 Flugzeuge*

Während die Ersteren mit Ontologien wie GermaNet behandelt werden können, brauchen wir für die Normalisierung der Letzteren weitere Informationen über die syntaktische und semantische Valenz von Wörtern, die von FrameNet-Daten zur Verfügung gestellt werden.

Die FrameNet-Annotation ist der Ausgangspunkt für zwei verschiedene Anwendungen: Erstens um automatisch IE-Templates zu erzeugen, zweitens um die FrameNet-Strukturen direkt in die Datenbank eines Question-Answering-Systems zu überführen.

II.1.2.2.2 FrameNet

FrameNet ist eine Datenbank, die syntaktische und semantische Valenzen dokumentiert, in Anlehnung an Charles Fillmores Theorie der thematischen Rollen ([Baker, Fillmore and Lowe 1998], siehe auch http://framenet.icsi.berkeley.edu/~framenet/). Semantisch verwandte Wörter werden anhand von Wortfeldern in eine hierarchische Struktur von Frames gruppiert.

Anstelle von universellen thematischen Rollen hat jeder Frame eine Menge von spezifischen Rollen (Frame-Elemente). Beispielsweise hat der Frame COMMERCE, mit Wörtern wie *kaufen, verkaufen* und *Verkauf* unter anderem die Frame-Elemente BUYER und SELLER.

So werden semantische Beziehungen zwischen Wörtern erfasst, die zu demselben Frame gehören und dieselben Frame-Elemente teilen. Frame-Argumente können als eine Bezeichnung der semantischen Argumente von Wörtern betrachtet werden, um sie vergleichbar zu machen. Daher werden mit der FrameNet-Annotation nicht nur Relationen wie Nahe-Synomyie erfasst, sondern auch inverse Relationen.

FrameNet wurde am ISCI in Berkeley entwickelt. Es hat im Englischen eine gute Abdeckung für verschiedene Domänen. Wir arbeiten an der Entwicklung einer deutschen Version von FrameNet, die Frame-Beschreibungen soweit wie möglich aus dem Englischen übernimmt. Ein großes deutsches Korpus wurde in Zusammenarbeit mit COLLATE mit FrameNet-Strukturen annotiert. Auf Basis eines kleinen hand-annotierten Korpus wurden mit den beschriebenen Techniken (semi-)automatische Annotationswerkzeuge und maschinelle Lernverfahren entwickelt.

II.1.2.2.3 Annotation von Text mit FrameNet

Die FrameNet-Annotation basiert auf syntako-semantischer Annotation von Texten. Wir verwenden die oben beschriebene PReDS-Annotation. Die FrameNet-Annotation besteht aus den folgenden Komponenten:

- Identifikation der Frame-evozierenden Wörter (lexikalische „Anker“)
- Identifikation des passenden Frames (entspricht in etwa Word Sense Disambiguation)
- Identifikation von Frame-Elementen
- Zusammenführen von partiell gefüllten Frames

Die ersten drei Aufgaben werden durch eine Menge von handkodierten Regeln behandelt. Da nicht alle Kombinationen von Regeln möglich sind, werden die Mengen von möglichen

Kombinationen berechnet. Aus diesen wird die Kombination mit der höchsten Gewichtung ausgewählt. Durch Anpassung der Regelgewichte kann dieser Prozess beeinflusst werden.

Die Information, die von den Regeln benutzt wird, enthält Informationen aus der FrameNet-Datenbank (Frame-evozierende Elemente, Valenzinformation) und die PReDS-Struktur (vergleichbar mit grammatischen Funktionen).

Die Regelmenge enthält einige allgemeine Regeln. Die meisten Regeln sind aber frame-spezifisch und lexikalisiert. Da diese Regeln die Verwendung eines Wortes in verschiedenen Domänen charakterisieren, ist das System nicht auf eine bestimmte Domäne eingeschränkt.

In der Schlussphase des Projektes wurde die Arbeit an den Transferregeln von PReDS zu FrameNet-Strukturen fortgesetzt. Ein besonderer Schwerpunkt lag dabei auf der Einbeziehung von Nicht-Core-Frame-Elementen. Neben den Core-Frame-Elementen, also für das jeweilige Frame zentralen thematischen Rollen, weist FrameNet auch Nicht-Core-Elemente aus, die nicht in allen Fällen realisiert sein müssen (z.B. TIME und LOCATION). Bei der Einbeziehung von Nicht-Core-Elementen in die Annotation ergibt sich eine höhere syntaktische Ambiguität. Die Transferregeln müssen daher zusätzliche Informationen (z.B. vom Named-Entity-Recognizer) berücksichtigen. In Zukunft sollen auch sortale Informationen (GermaNet) ergänzt werden.

II.1.2.2.4 Transfer in IE templates

Es wurden Transferregeln erstellt, die FrameNet-Strukturen in Templates für die Informationsextraktion abbilden. In einigen Fällen müssen dabei Unterscheidungen getroffen werden, die nicht in FrameNet repräsentiert werden. So wird beispielsweise der Frame COMMERCE nur dann in das Template für *Übernahme* abgebildet, wenn die GOODS eine Firma ist. Daher verwenden die Transferregeln Information von den vorigen Verarbeitungsschritten (Named-Entity-Erkennung) und sollen durch sortale Information (GermaNet) noch weiter verfeinert werden.

Zur Überführung der FrameNet-Strukturen in Informationsextraktions-Templates wurde ein Software-Modul entwickelt. Es basiert auf CFG-Regeln, die auf den erzeugten FrameNet-Strukturen matchen und daraus die entsprechenden IE-Templates extrahieren. Wir verwenden dabei die Templates, die von der Collate IE-Gruppe entwickelt wurden. Diese Übersetzung bildet die Grundlage für eine spätere Überführung in eine Datenbank, die für QA-Zugriffe auf Texte genutzt werden kann.

II.1.2.2.5 Transfer in eine QA-Datenbank

Die anspruchsvollste Anwendung auf Basis der FrameNet-Annotation ist der direkte Transfer der resultierenden Strukturen in eine Datenbank für QA-Anwendungen. Die Datenbank speichert Frames und Frame-Elemente sowie einen Verweis auf den Text. Anfragen werden vom System mit Hilfe spezieller Frage-Patterns in dieselbe FrameNet-Repräsentation überführt, wobei das Argument identifiziert wird, das die Antwort auf die Frage darstellt. Mit einer Datenbanksuche wird die korrekte Antwort gefunden und ein Ausschnitt des ursprünglichen Textes an den Benutzer zurückgegeben.

II.1.2.3 Modellierung von Informationszugriffsdialogen

Aus der Perspektive der Dialogmodellierung betrachten wir Question Answering als eine spezifische Art von Dialog, die in ähnlicher Weise wie andere Dialog-Anwendungen modelliert werden kann. Diese Perspektive unterscheidet sich von der im Information Retrieval üblichen Sichtweise, wonach Text Retrieval das Paradigma für Question Answering darstellt.

II.1.2.3.1 Information States in der Dialogmodellierung

Es wurde ein allgemeiner Ansatz für Dialogmodellierung aus den Projekten Trindi und Siridus übernommen, wo Dialoge durch Information States und Update-Regeln modelliert werden, die einen Informationszustand auf einen anderen abbilden. Wir verwenden die Entwicklungsumgebung TRINDIKIT, die die Definition von Information-State-Modellen unterstützt, und die Koordination mit externen Modulen wie Spracherkennung und Sprachsynthese ermöglicht.

Ein Information State bietet einen reichen Kontext für individuelle Dialogbeiträge oder Moves. Update-Regeln können spezifische Moves in eine Menge von primitiven Operationen auflösen, die allgemeine Dialogmodelle ermöglichen, in denen viele Details der Anwendung in Datenressourcen wie Lexika modularisiert sind. Wir haben ein solches Modell für Informations-zugriffsdialoge entwickelt, in dem die Details der Anwendung größtenteils in einer Planbibliothek ausgelagert sind. Im aktuellen Fall betreffen die Pläne die Beantwortung von Fragen aus einer Datenbank mit Wirtschaftsmeldungen.

II.1.2.3.2 Informationszugriffsdialoge

Dialoge für Informationszugriff bestehen aus zwei Phasen, die aus der Struktur der Anwendung folgen. Die erste Phase fokussiert auf die Frage, entscheidet, welcher Fragetyp involviert ist und stellt sicher, dass genug Informationen vorhanden sind, um eine Datenbank oder andere Informationsressource abzufragen. Dazu können auch weitere Informationen vom Benutzer angefragt werden. Das Dialogmodell ist flexibel genug, um jede Information auszunutzen, die spontan vom Benutzer geliefert wird oder im Kontext des Informationszustandes vorhanden ist. Die Abfrage kontextueller Information schließt die Auflösung von Anaphora und Ellipsen ein. Die zweite Phase des Dialogs folgt, sobald das System festgestellt hat, welche Information über die Frage vorhanden ist. Dann muss es die kooperativste Erwiderung auswählen. Diese kann entweder aus einer oder mehreren Antworten bestehen, oder aus einem weiteren Dialogschritt, um entweder die ursprüngliche Frage zu präzisieren oder die Menge der Antworten einzuschränken. Der Dialog kann auch nach einer Antwort des Systems fortgesetzt werden, wenn der Benutzer eine Präzisierung der Antwort verlangt oder eine weitere Frage stellt.

Ein besonderer Schwerpunkt für den dialogbasierten Informationszugriff sind ganz allgemeine Fragetypen, bei denen die Bedeutung und Funktion der Frage fast ausschließlich im Kontext liegt, wie z.B. *„Und was ist mit Ericsson“*, weil solche Fragen ganz deutlich und natürlich die direkten Vorteile des sprachbasierten Information Retrieval zeigen. Dialogstrategien zur Behandlung von solchen allgemeinen Fragetypen mit und ohne Kontext wurden entwickelt.

- Mit Kontext wird der intendierte Inhalt der Frage aus der vorhandenen Kontextinformation im Dialogmodell rekonstruiert. Weil dieser Dialogschritt zwangsläufig etwas unsicher bleibt, wird ein Bestätigungsmechanismus ausgelöst, damit der Benutzer eine mögliche Fehlinterpretation im rekonstruierten Inhalt noch rechtzeitig korrigieren kann.
- Ohne Kontext führen Fragen ohne deutlichen semantischen Inhalt direkt zu weiteren Klärungsfragen von der Seite des Systems.

Inzwischen werden auch Fragen nach der Begründung von Antworten durch das Anzeigen von Texten aus dem Korpus der Firmenmeldungen behandelt.

II.1.2.4 Bibliographie

Abney, Steven. 1996. “Partial Parsing via Finite-State Cascades.” *Proceedings of the ESSLLI '96 Robust Parsing Workshop.*

Becker, Markus, Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2002. "SProUT - shallow processing with typed feature structures and unification." *Proceedings of the International Conference on NLP (ICON 2002).*

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. "The Berkeley FrameNet Project". *Proceedings of the COLING-ACL.*

Bos, Johan, and Julia Heine. 2000. "Discourse and dialog semantics for translation." In Wolfgang Wahlster, Hg., *Verbmobil: Foundations of Speech-to-Speech Translation*.

Braun, Christian. 1999. *Flaches und robustes Parsen deutscher Satzgefüge.* Diplomarbeit, Universität des Saarlandes, Saarbrücken.

Elworthy, David. 2000. "Question Answering using a large NLP System". *Proceedings of the Ninth Text Retrieval Conference TREC 2000*

Harabagiu, Sanda M. and Maiorano, Steven J.. 1999. " Finding Answers in Large Collections of Texts: Paragraph Indexing + Abductive Inference". *Proceedings of AAAI Fall Symposium on Question Answering Systems.*

Hajičová, Eva. 2000. "Dependency-based underlying-structure tagging of a very large Czech corpus." *T.A.L.* **41**(1), pp. 47-66.

Larsson, S., P. Bohlin, J. Bos and D. Traum. 1999. *TRINDIKIT 1.0 Manual.* Trindi Report D2.2, University of Edinburgh, Scotland.

Larsson, Staffan. 2002. *Issue-based Dialogue Management.* PhD Thesis, Göteborg University.

Narayanan, Srini, Charles J. Fillmore, Collin F. Baker, and Miriam R. L. Petruck. 2002. "FrameNet Meets the Semantic Web: A DAML+OIL Frame Representation". *Proceedings of the 8th National Conference on AI.*

Pinkal, Manfred, C. J. Rupp, and Karsten Worm. 2000. "Robust semantic processing of spoken language." In:W. Wahlster, Hg., *Verbmobil: Foundations of Speech-to-Speech Translation*.

Peh, Li Shiuan and Ting, Christopher Hian Ann. 1996. "A Divide-and-Conquer-Strategy for Parsing". *Proceedings of the ACL/SIGPARSE 5th International Workshop on Parsing Technologies.*

Riloff, Ellen and R. Jones. 1999. "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping". *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99).*

Worm, Karsten L. 2000. *Robust Semantic Processing for Spoken Language.* Dissertation, Universität des Saarlandes, Saarbrücken.

II.1.3 Question Answering

Schwerpunkt der Forschung und Entwicklung waren Verbesserungen der Leistungsfähigkeit von umfangreichen Question-Answering-Systemen für offene Domänen. Ausgehend von dem QA-System Answerbus, das er an der Universität Michigan entwickelt hatte, hat Zhiping Zheng die Geschwindigkeit, Robustheit und Skalierbarkeit des Systems durch multi-level caching verbessert. Die Wissensbasis wurde erweitert um 8 Jahrgänge von CNN-Nachrichten (über 700.000 Artikel), die lokal in einer hocheffizienten Suchmaschine gespeichert sind, die in COLLATE entwickelt wurde. Dieselbe Suchmaschine und das QA-System wurden für ein Experiment eingesetzt, das in Zusammenarbeit mit der Firma XtraMind durchgeführt wurde. Darüber hinaus wurde in Zusammenarbeit mit dem Forschungszentrum Telekommunikation Wien eine gesprochensprachliche Schnittstelle für Question-Answering entwickelt.

II.1.3.1 AnswerBus

Answerbus ist ein web-basiertes multilinguales QA-System für offene Domänen. Das System wurde 2001 an der University of Michigan entwickelt, und in COLLATE weiterentwickelt. Die Effizienz des Systems wurde durch Caching auf zwei Ebenen stark verbessert. Der erste Cache enthält vor kurzer Zeit gestellte Fragen und deren Antworten. Die zweite Cache-Ebene enthält vorverarbeitete Web-Dokumente, wodurch wiederholte Zugriffe auf dieselbe Website und die wiederholte Analyse der Seiten entfallen können. Zur Zeit erhält das System 6000 bis 7000 Anfragen täglich.

Einige neue Techniken und Verfahren werden in AnswerBus verwendet. Wir haben einen neuen Algorithmus zum Vergleich von Fragen mit Sätzen entwickelt, mit dem entschieden werden kann, ob ein Satz eine mögliche Antwort auf eine Frage enthalten kann. Dieser Algorithmus wird eingesetzt, um mögliche Antworten aus einer großen Menge von Sätzen auszuwählen, und wird zum Filtern anstatt zum Sortieren der Antwortkandidaten eingesetzt. Abb. 8 zeigt die Architektur von AnswerBus.

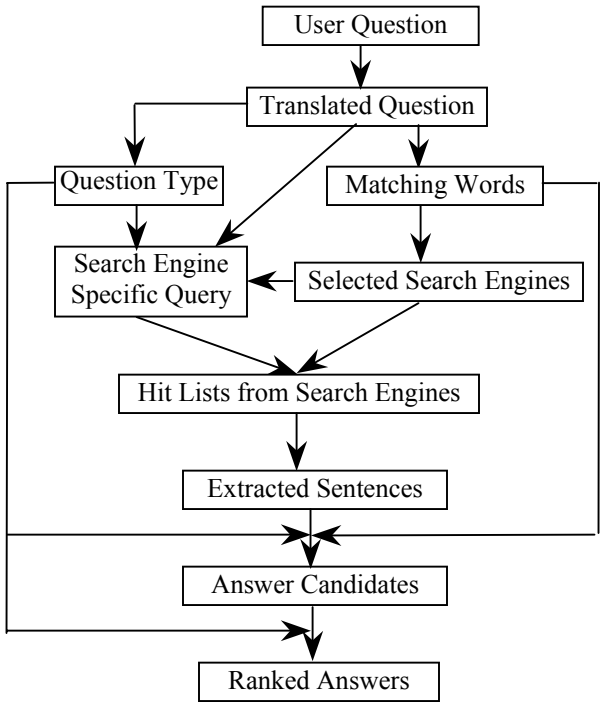

Abb. 8: Architektur das Question-Answering-Systems AnwerBus

Im Gegensatz zu anderen QA-Systemen wurde AnswerBus nicht speziell für die TREC-Wettbewerbe optimiert. Im Jahr 2002 wurde AnswerBus mit Fragen aus den TREC-Wettbewerben evaluiert. Anstatt der TREC-Dokumente wurde das WWW als Dokumentkollektion verwendet, aus welcher die Antworten extrahiert wurden. Daher sind die Ergebnisse (Tabelle 11) nicht direkt mit den veröffentlichen Ergebnissen aus den TREC-Wettbewerben vergleichbar.

Systems	Correct TOP 5	Correct TOP 1	NIST Score	T*max* (s)	T*min* (s)	T*mean* (s)	T*std dev*	L*mean* (byte)
AnswerBus	144	120	64.18%	15.06	3.79	7.20	3.07	141
IONAUT				44.88	2.78	12.51	6.81	1312
LCC	97	25	41.73%	342.52	4.30	44.24	32.63	178
QuASM	13	7	4.45%	284.29	2.61	20.72	33.92	1766
START	29	29	14.50%	62.07	2.02	9.84	7.45	

Tabelle 11: Evaluation von AnswerBus, im Vergleich mit TREC-Ergebnissen

II.1.3.2 AnswerBus News Engine

In diesem Experiment haben wir über 700.000 Nachrichten indexiert, die auf der CNN Website seit 1996 veröffentlicht wurden, und dazu ein effizientes Information-Retrieval-System entwickelt. Das Ziel des Experiments war es, die Techniken des AnswerBus-Systems mit einigen neuen Techniken, wie QA-spezifisischer Indexierung zu kombinieren, und ein QA-System für zeitsensitive Fragen aus der realen Welt zu entwickeln. Mit 700.000 Nachrichten haben wir unseres Wissens die weltweit größte Dokumentbasis für Question-Answering. Unser System ist für beliebig große Dokumentsammlungen skalierbar.

Aufgrund der lokalen Indexierung kann AnswerBus mögliche Antworten für eine Frage innerhalb von 2 bis 4 Sekunden aus der CNN-Sammlung extrahieren.

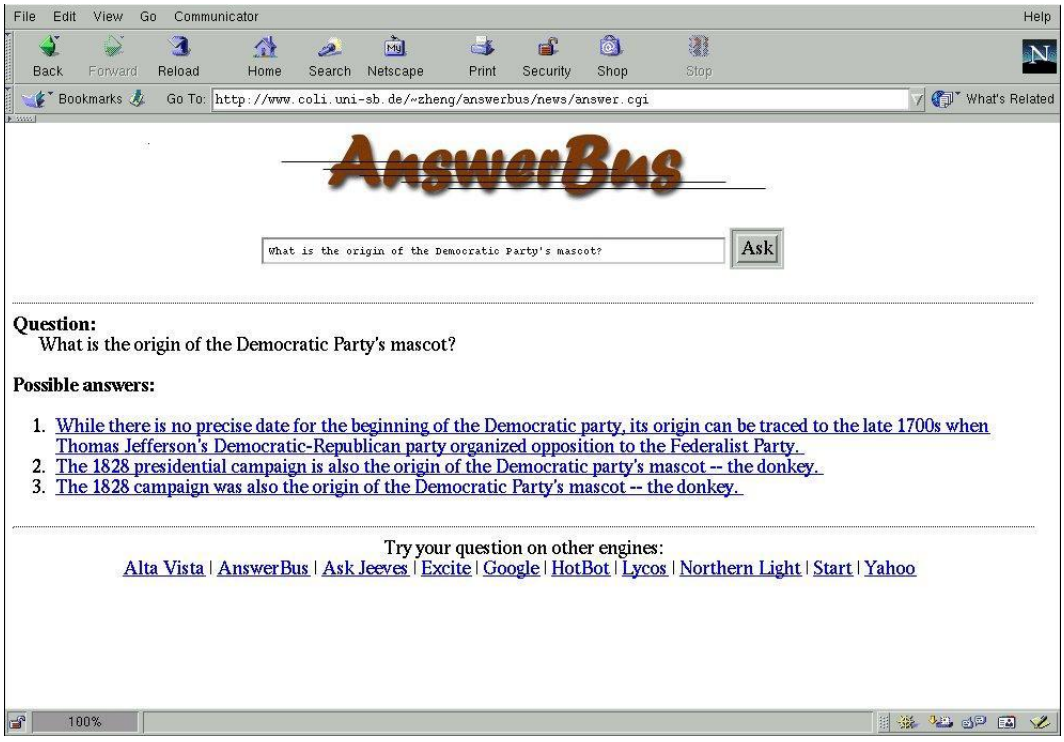

Abb. 9: AnswerBus Benutzerschnittstelle

II.1.3.3 Gesprochene Eingabe für Question-Answering

Gesprochene Eingaben für QA-Systeme eröffnen ein erhebliches Potential für das Anfragen von Informationen über Telefon und mit mobilen Kommunikationsgeräten. In Zusammenarbeit mit dem Forschungszentrum Telekommunikation Wien haben wir ein QA-System mit gesprochener Eingabe entwickelt. Für die Spracherkennung wurde das Diktiersystem Dragon

NaturallySpeaking 6.2, dessen Sprachmodelle für einen großen Korpus (280.000 Wörter) von Fragen adaptiert wurden. Tests mit einem männlichen englischen Muttersprachler und einer weiblichen Sprecherin wurden mit einem Headset-Mikrofon in einer ruhigen akustischen Umgebung durchgeführt.

Das System wurde mit 200 gesprochen Fragen aus der TREC 2002 QA-Aufgabe evaluiert. AnswerBus liefert kurze Ausschnitte aus Webseiten mit möglichen Antworten. Für jede der 200 Fragen erhielt AnswerBus zwei Eingaben, einmal den Wortlaut der Frage (verbatim typing), und dann die möglicherweise fehlerhafte Ausgabe des Spracherkenners (misrecognized speech). Die Ergebnisse sind in Tabelle 12 wiedergegeben.

	Speaker 1	Speaker 2
Misrecognized speech	39%	26%
Verbatim typing	58%	60%

	Speaker 1	Speaker 2
Degraded	12	34
Improved	5	0

Tabelle 12: Evaluation (accuracy) von gesprochenem Question-Answering

II.1.3.4 Automatische Textzusammenfassung

Wir haben am DUC2003-Wettbewerb für automatische Textzusammenfassung teilgenommen. Bei der Aufgabe 1 (sehr kurze Zusammenfassung) erzielte unser System dem dritten Platz in den Kriterien "Mean coverage" und "Median coverage" und den ersten Platz in den Kriterien “Mean length-adjusted coverage” und “Median length-adjusted coverage”.

Unser Ansatz beruht auf zwei Kriterien: 1) Wie wichtig ist jedes Wort im Text? und 2) Wie stark ist jedes Wort im Text anderen Wörtern im Text ähnlich?

Das erste Kriterium kann formuliert werden als "wie viel trägt ein Wort zum Thema das Textdokuments bei?" Nach unserer Auffassung unterscheidet sich die Häufigkeit eines Wortes in einem Textdokument von der Wichtigkeit des Wortes. Statt dessen ist die Häufigkeit des Wortes in einem großen Korpus von Dokumenten mit demselben Thema viel wichtiger. Daher kann bei der Zusammenfassung von Dokumenten mit einem bekannten Thema auf Worthäufigkeiten aus einem domänenspezifischen Korpus zurückgegriffen werden.

Das zweite Kriterium hilft zu entscheiden, welche Gruppe von Wörtern am wichtigsten für das Thema ist. Wenn zwei Wörter sehr verwandt sind, sollte wenigstens eines davon aus der Zusammenfassung entfernt werden. Auf Basis dieser Idee verwenden wir ein Programm, um zu ermitteln, wie ähnlich ein Wort mit anderen Wörtern ist. Wenn zwei Wörter in einem großen Korpus in ähnlichen Kontexten auftauchen, werden sie als distributionell ähnlich betrachtet. Das Programm erstellt eine Liste der Kontextwörter für jedes Wort im Korpus. Der TF*IDF-Wert jedes Wortes in der Liste wird berechnet, und Cosinus-Ähnlichkeit wird benutzt, um die Ähnlichkeit eines Wortes zu allen anderen Wortlisten zu berechnen.

II.1.3.5 Question-Answering für lokale Dokumentsammlungen

Dieses Experiment wurde in Zusammenarbeit mit XtraMind GmbH durchgeführt. Es wurden 50 Fragen verwendet, die Graesser's 16 Kategorien (Graesser et al. 1992) sowie 3 weitere Kategorien abdecken. Die Testergebnisse sind sehr positiv; die durchschnittliche Genauigkeit (accuracy) bei der besten Antwort ist 72%. Tabelle 13 stellt die Ergebnisse der Evaluation dar.

Question Type	Number	Top1	Top5	Wrong
1. Verification	3	1	1	1
2. Comparison	2	0	1	1
3. Disjunctive	2	2	0	0
4. Concept Completion	6	5	0	1
5. Definition	6	5	0	1
6. Example	3	3	0	0
7. Interpretation	3	2	1	0
8. Feature Specification	5	5	0	0
9. Quantification	6	4	0	2
10. Causal antecedent	3	2	0	1
11. Cause Consequence	0	0	0	0
12. Goal orientation	1	1	0	0
13. Enablement	0	0	0	0
14. Instrumental/Procedural	1	1	0	0
15. Expectational	1	1	0	0
16. Judgmental	3	1	0	0
17. Assertion	3	1	1	1
18. Request/Directive	0	0	0	0
19. Nils question	2	0	0	2

Tabelle 13: Evaluationsergebnisse für QA auf einer lokalen Dokumentsammlung

II.1.3.6 Bibliographie

John Burger, Claire Cardie, et al. (2002): Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST.

Graesser, A. C., Person, N. K., and Huber, J. D. (1992). Mechanisms that generate questions. In T. Lauer, E. Peacock, & A. C. Graesser (Eds.), Questions and information systems. Hillsdale, NJ: Lawrence Erlbaum Associates.

II.1.4 Web-Korpus: Struktur und Akquisition

Die Struktur eines Web-Korpus wurde definiert, und Werkzeuge und Verfahren für den Aufbau des Korpus implementiert. Der Web-Korpus soll einen großen Anteil der existierenden Webseiten für ein bestimmtes Wissensgebiet enthalten. Wir haben das Gebiet der Sprachtechnologie ausgewählt, da hierfür eine manuell annotierte Sammlung von 4700 ausgewählten Webseiten existiert, die am DKFI für das Informationszentrum LT-World gesammelt wurden. Zum Aufbau des Web-Korpus wird diese Dokumentenmenge expandiert durch das Verfolgen herein-kommender und herausgehender Hyperlinks, und durch die Suche nach ähnlichen Ressourcen.

Ein Webkorpus besteht aus einer Dokument-Datenbank und einer Hyperlink-Datenbank, so dass umfassende Informationen über die Dokumente und ihre Verbindungen durch Hyperlinks gespeichert werden. Die Struktur der Datenbanken ist im folgenden angegeben:

Feldname	**Inhalt / Beispiele**
URL	Lokation des Dokuments
Filetype / MIME type	Text/html, application/pdf, text/plain
Fulltext	Text des Dokuments, evtl. annotiert mit linguistischer Information wie POS, Named Entities, Phrasengrenzen
Fulltext Index	
Metadata	Author, language, date, category, keywords, abstract, type of page

Tabelle 14: Struktur der Dokumentdatenbank

Besondere Aufmerksamkeit haben wir dem Entwurf der Hyperlink-Datenbank gewidmet. Für jeden Hyperlink gibt es einen Eintrag in der Datenbank. Tabelle 15 zeigt die Struktur der Datenbankeinträge.

Unsere Hyperlink-Datenbank unterscheidet sich von früheren Arbeiten dadurch, dass wir für jeden Hyperlink eine reichhaltige Menge von Informationen sammeln (wie z.B. seine Position auf der Quellseite, oder den Text und Kontext des Quell-Ankers). Wir sammeln diese Information, da bestimmte Eigenschaften von Hyperlinks (z.B. Position am Anfang, Ende oder Mitte der Seite, oder Bezeichnung mit einem Text oder einem Bild) nützlich sein können um die Kategorie der Zielseite oder ihre Beziehung zu der Quellseite vorherzusagen.

Aus der Information in der Hyperlink-Datenbank können weitere Eigenschaften abgeleitet werden, die nützlich sein können, um Beziehungen zwischen Dokumenten zu entdecken. Zu diesem Eigenschaften gehören:

- Hyperlink innerhalb desselben Dokuments
- Hyperlink innerhalb desselben Servers vs. Hyperlink zu einem anderen Server
- Hyperlink innerhalb derselben second/third-level domain
- Hyperlink aufsteigend oder absteigend in der Verzeichnisstruktur
- Quellanker ist innerhalb einer Liste von Hyperlinks (Kontext `<OL>` oder `<UL>`)
- Navigationslink (mit Bezeichnungen *up, previous, next* ...)

Feldname	**Inhalt / Beispiele**
Source URL	Lokation des Quelldokuments
Target URL	Lokation of Zieldokuments
Source anchor type	Typ des Quell-Ankers (‘text’ oder ‘image’)
Source anchor text/context	Anker-Text/-Kontext, falls vorhanden
Source anchor position	Position des Quellankers im Quelldokument, relativ zur Dokumentlänge
Source anchor path	Position des Quellankers im Quelldokument, als Pfad von HTML/XML-Elementen von dem Root-Knoten
Target anchor position	Position des Zielankers im Zieldokument, relativ zur Dokumentlänge
Target anchor path	Position des Zielankers im Zieldokument, als Pfad von HTML/XML-Elementen von dem Root-Knoten
Target filetype	text/html, application/pdf, text/plain

Tabelle 15: Struktur der Hyperlink-Datenbank

II.1.4.1 Anwendung von Web-Korpora

Web-Korpora wurden für verschiedene Anwendungen verwendet, darunter Information Retrieval (Verwendung der Konnektivität zur Abschätzung der Qualität von Webseiten), Textkategorisierung (Verwendung der Konnektivität zur Vorhersage der Kategorie von Hyperlink-Zielen) und Textzusammenfassung unter Verwendung von Text im Quell-Anker. Wir beabsichtigen, die Konnektivitätsinformation auszunutzen, um Beziehungen zwischen Webseiten zu ermitteln. Unsere grundlegende Annahme ist, dass Links zwischen Webseiten Beziehungen zwischen den Objekten entsprechen, die durch die Webseiten beschrieben werden. Zum Beispiel zeigt ein Hyperlink zwischen der Homepage einer Person HP(P) und der Homepage einer Organisation HP(O) eine Beziehung zwischen der Person P und der Organisation O an. Eine genaue Analyse der Verbindungen zwischen Webseiten unterstützt daher die Entdeckung von Beziehungen in der realen Welt.

Die Webseiten in LT World bieten gute Trainings- und Testdaten, da die Kategorie und die Beziehungen manuell annotiert wurden.

II.1.4.2 Bibliographie

Peter Bailey, Nick Craswell and David Hawking. Engineering a multi-purpose test collection for Web retrieval experiments. *Information Processing and Management,* in press.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5)604-632, 1999.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems.* 30(1-7)107-117. 1998

Krishna Bharat, Andrei Broder, Monika Henzinger, Puneet Kumar and Suresh Venkatasubramanian. The Connectivity Server: fast access to linkage information on the Web. In: 7th International WWW Conference. Brisbane, Australia. 1998

Hans Uszkoreit and Brigitte Jörg. A Virtual Information Center for Language Technology: Ontology, Datastructure, Realization. In: *Nordic Language Technology Yearbook.* Museum Tusculanums Forlag, 2003.

II.1.5 Mobiles, multimodales und modulares Interface M3I

Das Framework M3I (Mobiles, multimodales, modulares Interface) ist ein Ergebnis des Projekts COLLATE. Es ist ein Framework für ressourcenadaptiven multimodalen Dialog mit mobilen Geräten. Die Bandbreite der Anwendungen umfasst Systeme, die auf tragbaren Geräten wie PDAs oder Smartphones laufen; im Speziellen wurden ein Fußgängernavigationssystem, ein Shopping-Assistent und ein Museumsführer bearbeitet. M3I stellt den Systemen eine Schnittstelle für multimodale Eingabe (gesprochene Sprache und Stift) und Ausgabe (gesprochene Sprache und Grafik) zur Verfügung.

Der Begriff *Ressourcenbeschränkung* bezieht sich sowohl auf technische wie auf kognitive Ressourcen. Für die Anwendungen wird die folgende technische Umgebung vorausgesetzt: die Geräte haben ein kleines Display, beschränkte Speicherkapazität und geringe Rechenleistung. Eine Breitbandverbindung mit einem Server ist in der Regel verfügbar (was realistisch ist, da die Abdeckung von Breitbandnetzwerken wie WLAN und UMTS ständig erweitert wird). Es muss jedoch mit zeitweisen Ausfällen der Netzwerkverbindung gerechnet werden. In diesem Falle muss das System ohne Netzwerk weiter funktionieren, wenn auch mit eingeschränktem Funktionsumfang und geringerer Interaktionsqualität.

In Bezug auf kognitive Ressourcenbeschränkungen werden Erwachsene mittleren Alters und ältere Menschen als beispielhafte Benutzergruppen betrachtet. Ältere Menschen gehören zu den Gruppen, die zuletzt vom Zugang zu Computern profitieren. Sie haben Schwierigkeiten beim

Umfang mit Technologie, da sie oft an kognitiven Beschränkungen leiden, wie z.B. altersbedingte degenerative Prozesse, motorische Einschränkungen, Probleme des Kurzzeitgedächtnisses sowie verminderte visuelle und auditive Fähigkeiten (Jorge, 2001). Diese Einschränkungen werden oft verstärkt durch Unkenntnis der Technologien und verschiedene individuelle Lernkurven.

M3I stellt die folgenden Dienste zur Verfügung: Spracherkennung und (interne) Gesteninterpretation, Akquisition von Benutzermodellen, Auswahl von passenden Adaptationsstrategien, und die Generierung von passenden Ausgaben. Die Clients (mobile Geräte) beinhalten Basisversionen dieser Dienste für den Fall, wo keine Netzwerkverbindung vorhanden ist. Der Server stellt umfassende Versionen der Dienste zur Verfügung, z.B. parallele Spracherkennung mit größerem Vokabular und Sortierung der Ergebnisse.

II.1.5.1 Architektur

Die oben genannten Annahmen haben Auswirkungen auf die Systemarchitektur und -funktionalität:

a) eine Netzwerkverbindung zwischen verschiedenen Applikationen und einem zentralen Server müssen bereitgestellt werden, und eine Kommunikationskomponente für heterogene Systeme ist erforderlich,
b) die Architektur sollte die einfache Integration neuer Dienste ermöglichen,
c) für jeden Dienst soll die Server-Version zuerst implementiert werden, und - wenn die Probleme wohlverstanden sind - kann eine vereinfachte Version abgeleitet und für Client-Systeme implementiert werden.

Folglich wurde der M3I-Server in Java implementiert, um das schnelle Erstellen von Prototypen zu erleichtern, und embedded C++ wurde für die Clients gewählt, um optimale Performance zu gewährleisten.

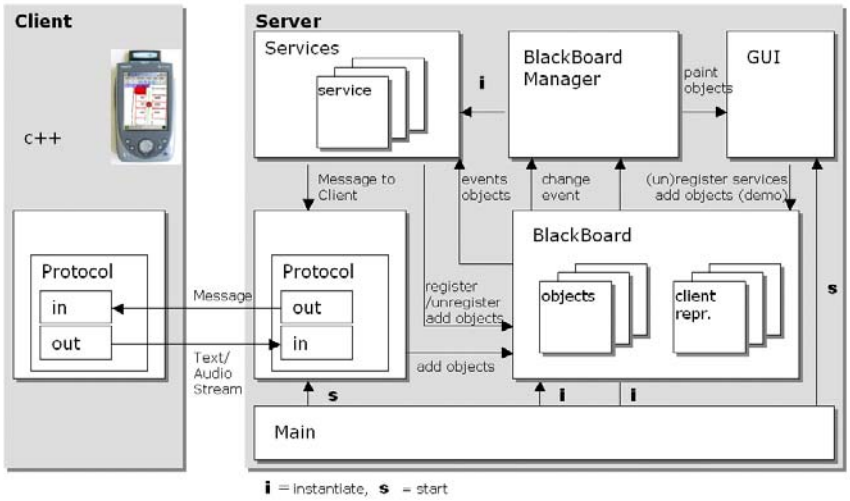

Abb. 10: Die M3I Architektur

Abb. 10 zeigt die M3I Client/Server-Architektur. Clients, mobile Geräte wie HP IPQA, kommunizieren mit dem Server durch eine Socket-Verbindung. Clients können über ein spezielles Übertragungsprotokoll Textinformationen (ASCII) oder gesprochene Sprache (byte streams) zum Server schicken. Das zentrale Element des Servers ist ein Blackboard, wo Informationen als Java-Objekte gespeichert werden.

Dienste können sich für Blackboard-Objekte eines bestimmten Typs registrieren. Immer wenn ein neues Objekt hinzugefügt oder ein existierendes Objekt geändert wird, wird ein entsprechendes

Ereignis generiert und die registrierten Dienste erhalten das neue bzw. aktualisierte Objekt. Die Ausgabe des Dienstes wird als neues Objekt auf dem Blackboard repräsentiert, das wiederum andere Dienste aktiviert.

Dieses Verfahren wird auch für die Akquisition der Benutzermodelle (Geschlechts- und Alterserkennung) eingesetzt. Wenn der Benutzer eine gesprochene Eingabe liefert, wird die Sprache auf dem mobilen Gerät erfasst und in Echtzeit zum Server übertragen, der ein entsprechendes Objekt für ein Audio-Objekt auf dem Blackboard erzeugt. Der Dienst *voice feature extractor,* der sich für Audio-Objekte registriert hat, extrahiert Merkmale aus dem Sprachsignal und schreibt ein neues Objekt vom Typ *feature object* auf das Blackboard. Daraufhin wird der *classifier service* aktiv, der eine Implementation des maschinellen Lernverfahrens ist, das in Abschnitt II.1.5.2 beschrieben wird. Der Klassifikator schreibt ein neues Objekt vom Typ *user model object* auf das Blackboard. Der Dienst *adaptive output service,* der sich für diese Art von Objekten registriert hat, schickt eine Nachricht an den Client, der eine Ausgabe erzeugt, die für den aktuellen Benutzer angepasst ist.

II.1.5.2 Akquisition von Benutzermodellen durch Abschätzung von Alter und Geschlecht

Bei der Anpassung des Systemverhaltens an die Bedürfnisse älterer Menschen ergeben sich zwei Fragestellungen.

1. Welche Anpassungen soll das System vornehmen, wenn es weiß, dass der Benutzer zu der Gruppe älterer Menschen gehört?
2. Wie kann das System dieses Wissen erlangen?

In Bezug auf die erste Frage schlagen Müller und Wasinger (2002) vor, dass in einem multimodalen Dialogsystem die Sprachausgabe langsamer/lauter sein sollte, und dass graphische Elemente wie Toolbars, Buttons, Karten und Texte größer dargestellt werden. In COLLATE lag der Schwerpunkt der Forschung auf der zweiten Frage, wobei gesprochene Sprache als eine wichtige und ergiebige Quelle für Informationen über den Sprecher identifiziert wurde. Wenn wir eine Stimme hören, können wir in den meisten Fällen das Geschlecht des Sprechers erkennen, das Alter abschätzen und vielleicht auch die Stimmung des Sprechers in Bezug auf Stress oder Emotionen erkennen.

II.1.5.2.1 Extraktion von Merkmalen aus dem Sprachsignal

Wir betrachten akustische und prosodische Merkmale, die relativ einfach vor dem Spracherkennungsprozess extrahiert werden können. Für die Erkennung von Alter und Geschlecht werden die Merkmale *jitter* (maximale Perturbation der Grundfrequenz) und *shimmer* (Variation der Spitzenamplitude) extrahiert. Die Extraktoren wurden mit dem frei verfügbaren phonetischen Analyseprogramm PRAAT implementiert, das fünf Jitter-Algorithmen und drei Shimmer-Algorithmen bereitstellt.

II.1.5.2.2 Klassifikationsverfahren

Zu Anfang wurde ein Experiment durchgeführt, um herauszufinden, ob es überhaupt möglich ist, Benutzer aufgrund des Sprachsignals nach Geschlecht und Alter zu klassifizieren. In dem Experiment wurden die gebräuchlichsten maschinellen Lernverfahren verglichen.

Die Testdaten bestanden aus drei Korpora von Sprachsignalen, die mit Alter und Geschlecht des Sprechers annotiert waren: ein Korpus von Scansoft, der TIMIT-Korpus und ein Korpus der im M3I-Projekt gesammelt wurde. Tabelle 16 zeigt die Verteilung von Sprechern.

corpus	speakers > 60	speakers < 60	female speakers	male speakers	total

SCANSOFT	347	5	222	130	352
TIMIT	5	625	192	438	630
M3I	0	46	28	18	46
total	352	676	442	586	1028

Tabelle 16: Verteilung der Sprecher in den Testdaten

Die folgenden Klassifikationsmethoden (und Parameter) wurden eingesetzt:: C4.5 decision tree induction (DT), artificial neural networks (ANN, learning rate 0.15, momentum 0.2, 500 iterations), k-nearest neighbors (kNN, k=5, simple distance weighting), naive Bayes, (NB) und support vector machines (SVM, C=20, polynomial kernel with degree 4). Wir benutzten die WEKA machine learning tools (Witten and Frank, 1999).

In diesem Experiment wurden die Durchschnittswerte der acht Jitter- and Shimmer-Merkmale jeder Person verwendet.

Tabelle 17 zeigt die Genauigkeit der Klassifikation. Als Baseline (BL) geben wir einen Klassifikator an, die immer die am häufigsten auftretende Klasse voraussagt, d.h. ältere Menschen (88%) and männlich (59%).

	C5	ANN	KNN	NB	SVM	BL
Geschlecht	69,10	81,73	76,41	67,26	70,43	58,78
Alter	92,41	96,75	95,76	91,25	96,45	88,30

Tabelle 17: Vorhersagegenauigkeit verschiedener maschineller Lernverfahren

Die Ergebnisse zeigen, dass es möglich ist, Altersgruppe und Geschlecht von Sprechern anhand akustischer Merkmale vorherzusagen, wobei ANNs die beste Genauigkeit zeigten.

In Tabelle 18 wird die Genauigkeit des ANN-Klassifikators für die verschiedenen Klassen illustriert. Die erste Zeile zeigt die Ergebnisse mit einem unbalancierten Datenmenge, und die zweite die Verbesserungen, die sich durch die Einbeziehung des TIMIT-Korpus ergaben.

	non-elderly	elderly	female	male
unbalanced set	83%	98%	70%	89%
balanced	93%	98%	90%	93%

Tabelle 18: Anteil der korrekten Voraussagen (ANN)

II.1.5.2.3 Zweistufiges maschinelles Lernverfahren

Wir haben ein zweistufiges maschinelles Lernverfahren zur Erkennung von Alter und Geschlecht entwickelt. Die Klassifikatoren der unteren Ebene behandeln Sensordaten (Jitter und Shimmer). Die Verlässlichkeit dieser Klassifikatoren hängt vom Kontext und von Umgebungseinflüssen wie Hintergrundgeräuschen und Mikrophoncharakteristika ab. Darüber hinaus existieren Abhängigkeiten zwischen den Ergebnissen der verschiedenen Klassifikatoren der unteren Ebene. In Bezug auf unsere Anwendung ist wichtig, dass die Stimmen von Frauen und Männern unterschiedlich altern (Linville, 2001). Wir berücksichtigen diese Abhängigkeiten mit Hilfe Bayesscher Netzwerke (Pearl, 1988) auf der oberen Ebene zur Kombination der Ergebnisse und zur Repräsentation der kausalen Abhängigkeiten zwischen der Klassifikationsqualität und Umgebungseinflüssen.

Abb. 11 zeigt die Struktur des Bayesschen Netzwerks. Es gibt direkte kausale Abhängigkeiten zwischen dem Alter/Geschlecht des Benutzers und dem Ergebnis des Klassifikators für ein bestimmtes Sprachsignal.

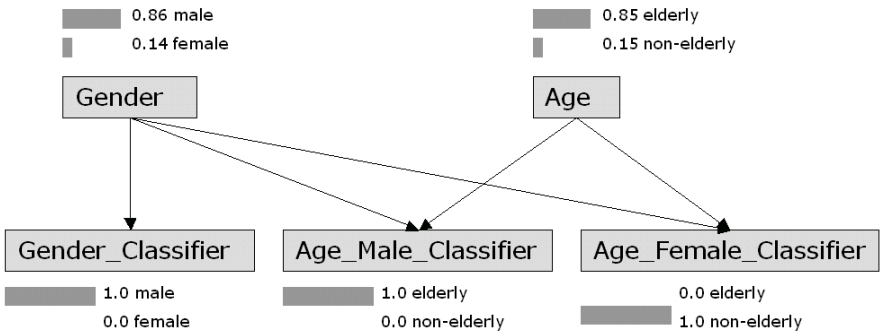

Abb. 11: Bayessche Netzwerke zur Integration der Klassifikationsergebnisse

Die Verwendung Bayesscher Netzwerke zur Integration der Klassifikationsergebnisse hat mehrere Vorteile:

a) sie liefert explizite Wahrscheinlichkeitswerte für die möglichen Hypothesen,
b) sie verbessert die Qualität der Ergebnisse im Vergleich zu einer Klassifikationshierarchie mit sequentiellen Entscheidungen (zuerst Geschlecht, dann Alter), indem sie die Unsicherheit bei der Entscheidung über das Geschlecht in das Bayessche Entscheidungsverfahren einbezieht, und
c) Konfidenzwerte der Klassifikatoren der unteren Ebene können in einem Bayesschen Netzwerk berücksichtigt werden.

II.1.5.3 Medienfusion in einer M3I-Plattform

Multimodale Interaktion in mobilen Umgebungen ist eine Herausforderung für die Forschung. Das liegt in ersten Linie an der eingeschränkten Leistungsfähigkeit von PDAs und Smartphones, an den dynamischen Umgebungen (im Gegensatz zu stationären Geräten) und dem Wechsel zwischen Umgebungen im Freien und in Gebäuden.

Medienfusion ist die Fusion von verschiedenen Benutzereingaben wie Sprache und Gesten zu einer einzigen eindeutigen und modalitätsunabhängigen Repräsentation. Das Ziel ist es, modalitätsspezifische Eingaben des Benutzers mit verschieden Sensoren wie Mikrofonen und Touchscreens in eine einzige Repräsentation zu überführen, die unabhängig von den Modalitäten ist. Diese Fusion ist wichtig, um anderen Programmmodule eine einzige Interpretation der Benutzereingabe zu liefern. Medienfusion ermöglicht eine natürliche und flexiblere Mensch-Maschine-Interaktion, und trägt wesentlich zu einer Steigerung der Robustheit bei. Eine herausragende Eigenschaft dieser Implementation ist, dass alle multimodalen Eingaben in Echtzeit auf dem mobilen Endgerät (z.B. Pocket PC) verarbeitet werden. Eine weitere Besonderheit ist die Verwendung eines gemeinsamen Blackboards, wodurch das Modul für verschiedenartige Eingaben erweitert werden kann. Neben gesprochener Sprache und Gestik können diese Eingaben von verschiedenen Sensoren oder einer Kamera kommen.

II.1.5.3.1 Szenarien

Das Fußgängernavigations-System ermöglicht einem Benutzer, vordefinierte Routen auf ein Mobilgerät zu übertragen, und eine Route für Navigation und Exploration im Innen- und Außenraum auszuwählen. Der Navigationsmodus führt den Benutzer mittels einer Kombination von Sprache und Graphik zu seinem Zielpunkt, während der Explorationsmodus es Benutzern ermöglicht, Informationen über Objekte in ihrer Umgebung abzufragen. Anfragen können mittels Spracheingabe, Gesten oder einer Kombination der beiden gestellt werden.

Das in C/C++ implementierte System enthält eine formantbasierte Sprachsynthese und eine Spracherkennung mit dynamischen Grammatiken (IBM Embedded ViaVoice). Die 2D/3D-Graphiken werden mit dem Parallel Graphics Cortona VRML Browser erzeugt, der auch die Gesten interpretiert. Weiterhin dient ein GPS-Modul der Lokalisierung im Außenraum und ein Infrarotmodul der Lokaliseriung im Innenraum.

Als weiteres Szenario wurde in der letzten Projektphase ShopAssist implementiert, wobei Fortschritte in den Bereichen GUI-Design, Sprachverarbeitung, Graphik und Client-Server-Kommunkation erzielt wurden.

II.1.5.3.2 Medienfusionskomponente

II.1.5.3.2.1 Architektur

Abb. 12 zeigt die drei wesentlichen Eingabekomponenten für die Medienfusionskomponente: die eingebetteten Erkenner und Sensoren (oben), die Wissensquellen und dynamischen Grammatiken (rechts) und die Benutzermodell-/History-Information (links). Die eingebettete Spracherkennung und Sprachsynthese, die Verarbeitung von Stift-Gesten auf dem Bildschirm des Pocket-PCs (Intra-Gesten), einige Sensorinformationen, dynamisch generierte Grammatiken und die Wissenskomponente für die Medienfusion sind alle erfolgreich integriert worden.

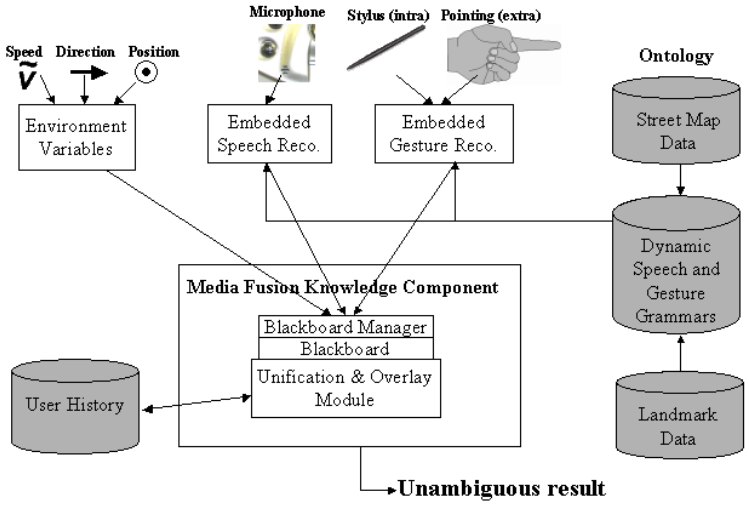

Abb. 12: Integration der Medienfusionskomponente (MF) in das Fußgängernavigationssystem.

Die Abfolge der Schritte für die Medienfusion ist in Abb. 13 illustriert.

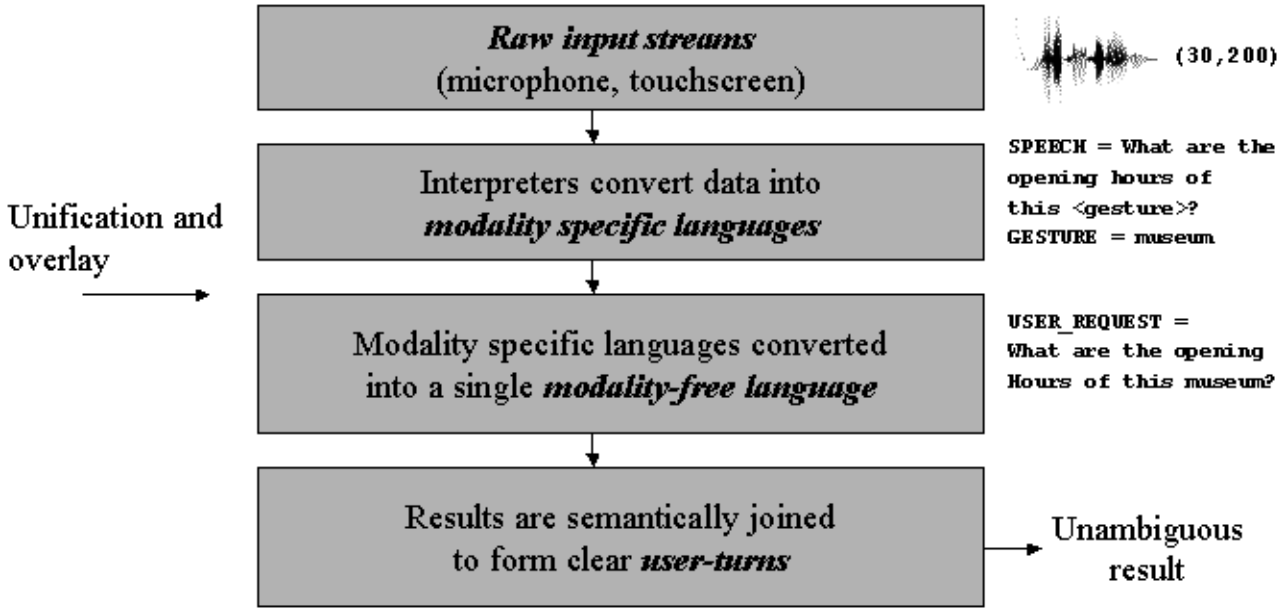

Abb. 13: Prozess der Erzeugung einer eindeutigen Repräsentation aus verschiedenen Informationsquellen.

Eine wesentliche Komponente beim Entwurf des Medienfusionsmoduls ist das zentrale Blackboard und der Blackboard-Manager (Abb. 14). Information auf dem Blackboard erhält einen Medientyp (z.B. Sprachsignal, Geste, Sensor) und eine Zeitmarke. Das Unifikations- und Overlay-Modul ist dann verantwortlich für die tatsächliche Fusion der Daten.

Source Type	MI Type	MI	Unique ID	Time	Conf. Value	Modality Ptr. or Value
PDA	Sensor	Direction	100	1046788415000	--	directionVal
PDA	Sensor	Velocity	101	1046788415000	--	velocityVal
PDA	Speech	Microphone	102	1046788415000	0.8	*speechPtr
PDA	Gesture	Stylus	103	1046788415500	0.9	*objectPtr

Abb. 14: Beispiel der Anwendung des Blackboards für Medienfusion.

In der letzten Projektphase wurde ein theoretisches Design eines Medienfusionsalgorithmus erarbeitet, der Bayessche Netze (BN) benutzt.

II.1.5.3.2.2 Benutzerinteraktion

Die Interaktionsmodalitäten für den Benutzer sind aktuell gesprochene Sprache und Gesten. Es gibt zwei Arten von Gesten: Intra-Gesten (Zeigen auf Objekte auf dem mobilen Gerät) und Extra-Gesten (Zeigen auf Objekte in der realen Welt). Intra-Gesten werden mit einem Stift ausgeführt, während Extra-Gesten Sensorinformationen wie die aktuelle Position und Orientierung des Mobilgeräts sowie die Geschwindigkeit benötigen, die durch einen Accelerometer und eine GPS-Einheit erfasst werden. Vermittels dieser Sensorinformationen kann der Benutzer auf Objekte in der Welt zeigen während er spricht, wie in Abb. 15 illustriert.

Kombination von Sprache und Intra-Geste; Zeige- und Gleit-Gesten

Kombination von Sprache und Extra-Geste.

Abb. 15: Kombinationen von Gesten

Das System unterstützt drei verschiedene Interaktionsmethoden: nur Sprache, nur Gesten und die Kombination von Sprache und Gesten, die der Schwerpunkt unserer Arbeit sind. Beispiele sind die Eingaben “what is this <gesture>?”, “take me to here <gesture>”, und ”tell me about the church over there <gesture>”. In allen diesen Beispielen kann der Benutzer auf Objekte zeigen, die er sehen, aber nicht benennen kann. Durch Einsatz von Wissensquellen und modalitätsspezifischen Grammatiken kann so eine hohe Flexibilität der Interaktion erzielt werden.

II.1.5.3.2.3 Grammatiken und Repräsentationssprachen

Jede Modalität benötigt ihre eigene Menge von Grammatiken, um die Benutzereingaben zu analysieren. Diese Grammatiken existieren für verschiedene Verarbeitungsebenen, und bilden gemeinsam eine Interaktionsbibliothek. Auf einer höheren Ebene können Gesten in Intra- und Extragesten klassifiziert werden, während sie auf einer niedrigeren Ebene (Abb. 16 links) klassifiziert werden als Zeigegesten (z.B. zur Identifikation von Gebäuden) und Gleitgesten (z.B. zur Identifikation von Straßen). Diese Interaktionsbibliotheken ermöglichen eine reichere Mensch-Maschine-Interaktion. Abb. 16 zeigt rechts verschiedene Sprachgrammatiken. Diese Information über Landmarken und Straßennahmen wird aus der Wissensbasis extrahiert und in einem XML-Format repräsentiert.

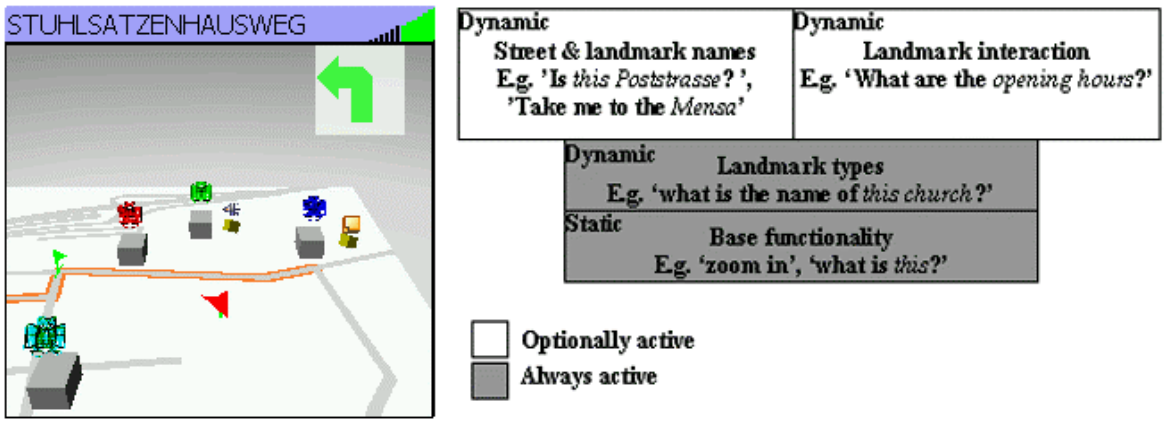

Abb. 16: Statische und dynamische Sprachgrammatiken, und PDA-Display mit einer Route mit Landmarken als Interaktionsobjekten..

II.1.5.4 Bibliographie

[Baken and Orlikoff, 2000] R. Baken and R. Orlikoff, *Clinical measurement of speech and voice (2 nd edition)*. San Diego: Singular publishing Group, 2000.

[Dagum, 1992] P. Dagum, A. Galper, and E. Horvitz, "Dynamic network models for forecasting," in *Uncertainty in Artificial Intelligence:Proceedings of the Eight Conference*. San Francisco: Morgan Kaufmann, 1992, pp. 41–48.

[Jorge, 2001] J.Jorge. Adaptive tools for the elderly, new devices to cope with age-induced cognitive disabilities. *In Proceedings of the EV/NSF Workshop in Universal Accessibility of Ubiquitous Computing*: Providing for the Elderly.

[Linville, 2001] S. E. Linville, Vocal Aging. San Diego, Ca: Singular, 2001.

[Pearl, 1988] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[Wasinger, 2003] R.Wasinger, C.Stahl and A.Krüger. *M3I in a Pedestrian Navigation & Exploration System*. In: Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices.

[Witten and Frank, 1999] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations*. Morgan Kaufmann Publishers, 1999.

II.2 Voraussichtlicher Nutzen

Die zwei Hauptziele des Projekts COLLATE-UdS sind

- die Entwicklung, Verbesserung und Optimierung der Interaktion mit Informationssystemen durch Question Answering, Dialog, Media Fusion und Benutzeradaptation,
- die Weiterentwicklung von Technologien zum Wissenserwerb aus Texten durch Verfahren der Informationsextraktion.

In diesen beiden großen Bereichen wurden zahlreiche Ergebnisse erzielt, die in praktischen Anwendungen verwertet werden können. Die Verwertung erfolgt durch Transfer von Wissen und Technologien durch Publikationen, Zusammenarbeit mit Firmen, Vergabe von Lizenzen, und Ausbildung von Wissenschaftlern und Studenten. Es bestehen Kontakte mit mehreren Firmen, die an einer Verwertung der Ergebnisse interessiert sind.

Wirtschaftlicher Nutzen

Im Bereich Question Answering bestehen gute Verwertungsmöglichkeiten im Bereich der nächsten Generation von Suchmaschinen, und im Wissensmanagement. Es bestehen zur Zeit Kontakte mit mehreren interessierten Firmen. Daneben wurde im Zusammenhang mit der bevorstehenden Gründung eines Spinoff-Unternehmens die Marktsituation für Question-Answering-Technologien untersucht und ein Business-Plan erarbeitet.

Im Bereich der Benutzerschnittstellen (Media Fusion und multimodaler Dialog, benutzeradaptierte Spracherkennung), gibt es ausgezeichnete Verwertungsmöglichkeiten durch Transfer von Technologien in die Industrie, insbesondere im Sektor Mobilkommunikation (Telekom-Firmen, Gerätehersteller, Automobilindustrie), wo multimodale Schnittstellen mit Sprachein- und –ausgabe zunehmende Bedeutung gewinnen. Hier gibt es zahlreiche konkrete Kontakte und Zusammenarbeit mit Anwenderfirmen.

Im Bereich der Informationsextraktion wurde die multilinguale Technologieplattform und Entwicklungsumgebung SProUT im Projekt entwickelt und Grammatiken für sieben Sprachen erstellt und getestet. An dieser Technologie besteht großer Bedarf in den Sektoren Wissensmanagement, Dokumentverwaltung und Suchmaschinen. SProUT ist schon bei mehreren Firmen im Einsatz. Transfer und Verwertung erfolgen hier durch Lizenzen und Industrieprojekte.

Wissenschaftliche und/oder technische Bedeutung

Die im Projekt entwickelten Verfahren und Technologien sind auf höchstem wissenschaftlich/technischem Niveau. Durch Publikationen in internationalen Konferenzen und Teilnahme an internationalen Evaluationen (Text Retrieval Conference TREC, Document Understanding Conference DUC) wird die Vergleichbarkeit mit den weltweit führenden Systemen sichergestellt. Die entwickelten Technologien stellen eine gute Basis für anwendungs- und grundlagenorientierte Nachfolgeprojekte dar. Die im Projekt aufgebauten Sprachressourcen (annotierte Korpora) und Evaluationswerkzeuge bilden die Grundlage für eine vergleichende Evaluation der Ergebnisse, und werden bereits in verschiedenen Projekten genutzt.

Wissenschaftliche und wirtschaftliche Anschlussfähigkeit

Robuste, präzise und multilinguale Informationsextraktion ist eine der Schlüsseltechnologien zur Bewältigung der Informationsflut durch die ständig zunehmende Anzahl elektronisch verfügbarer Dokumente, und stellt somit ein wissenschaftlich wie wirtschaftlich hochinteressantes Gebiet dar. Durch ihre erfolgreiche Implementierung im SProUT-System werden wissenschaftliche Ergebnisse direkt in praktisch anwendbare Software umgesetzt. Für darauf aufbauende Systeme und Produkte besteht bereits jetzt wirtschaftliches Interesse. Es bestehen vielfältige Anwendungsmöglichkeiten in anderen Gebieten, die einer robusten, effizienten und adaptiven Sprachtechnologie bedürfen. Beispielhaft seien hier Text Data Mining, komplexe Informationsmanagementsysteme, multimediale Information-Retrieval-Systeme, oder sprachbasierte E-Commerce- und Agentenarchitekturen genannt. Sowohl in wissenschaftlicher wie auch in wirtschaftlicher Hinsicht bestehen hervorragende Aussichten für einen erfolgreichen Transfer und anwendungs- und grundlagenorientierte Nachfolgeprojekte.

Im Bereich der Interaktion mit Informationssystemen durch Question Answering, Dialog, Media Fusion und Benutzeradaptation besteht sehr großer wirtschaftlicher Bedarf, vor allem im Bereich der mobilen Kommunikationssysteme. Dadurch bestehen hier ausgezeichnete Möglichkeiten für erfolgreichen Transfer und Anschlussprojekte.

II.3 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Durch die Teilnahme an internationalen Konferenzen und Evaluationen besteht ein guter Kontakt zu anderen Forschungsgruppen, die auf dem Gebieten des Vorhabens arbeiten. Zugleich wurde

durch die Annahme von Publikationen bei den führenden internationalen Konferenzen, Workshops und Journals das internationale Niveau und die Akzeptanz der Projektergebnisse bestätigt.

Im Bereich der Informationsextraktion existiert in den USA das Förder- und Evaluationsprogramm ACE (Automatic Content Extraction), an dem allerdings nur inländische Gruppen teilnehmen dürfen. Eine offene Evaluation für multilinguale Informationsextraktion wurde im Rahmen der CONLL-Konferenz (Computational Natural Language Learning) durchgeführt; die CONLL-Testdaten stehen dem Projekt COLLATE jetzt zur Verfügung, und ermöglichen den Vergleich mit anderen Gruppen. Im Bereich Question Answering ermöglichen das amerikanische Evaluationsprogramm TREC-QA mit englischen Testdaten und das europäische Programm CLEF-QA, bei dem Wissenschaftler aus COLLATE eine führende Rolle spielen, den Vergleich mit anderen internationalen Gruppen. Im Bereich der automatischen Textzusammenfassung besteht ein Vergleich mit anderen Gruppen im Rahmen des Evaluationsprogramms DUC (Document Understanding Conference), bei dem die in COLLATE entwickelten Systeme sehr gut abgeschnitten haben. Die an COLLATE beteiligten Arbeitsgruppen wurden zu weiteren Verbundprojekten eingeladen, in denen sie mit den besten europäischen und internationalen Gruppen zusammenarbeiten. Auf dem Gebiet der Dialogsysteme wurde im 6. Rahmenprogramm der EU das integrierte Projekt TALK bewilligt, in das die Ergebnisse von COLLATE einfließen. Im Projekt COLLATE ist es uns gelungen, Technologien und Ressourcen für die deutsche Sprache (Informationsextraktion, FrameNet) auf internationales Niveau zu bringen, international anerkannte Spitzenforschung zu leisten (M3I, Textzusammenfassung, Dialog), und weltweit genutzte Dienste und Ressourcen (Question Answering) anzubieten.

II.4 Erfolgte und geplante Veröffentlichungen des Ergebnisses

Veröffentlichungen

Avgustinova, T. (2002). “Shared Grammatical Resources for Slavic Languages”. Selected topic in multilingual grammar design with special reference to Slavic morphosyntax. Habilitationsschrift. Universität des Saarlandes.

Avgustinova, T. (2003) Metagrammar of systematic relations: a study with special reference to Slavic morphosyntax. Syntactic Structures and Morphological Information. Uwe Junghanns and Lika Szucsich (Editors), Mouton de Gruyter, Berlin / New York 2003. (Interface Explorations 7) ISBN 3-11-017824-9. Pages 1-24.

Avgustinova, Tania and Hans Uszkoreit (2003). Towards a typology of agreement phenomena. The Role of Agreement in Natural Language: TLS 5 Proceedings. W. E. Griffin, ed. Austin, TX, Texas Linguistics Forum. 53: 167-180.

Baumann, S., Klüter, A. and Norlien, M. (2002). „Using natural language input and audio analysis for a human-oriented MIR system”. In: *Proceedings of WEDELMUSIC, the International Conference on Web Delivering of Music*. Darmstadt.

Baus, J., Krüger, A. and Wahlster, W. (2002). „A Resource-Adaptive Mobile Navigation System”. In: *Proceedings of the 2002 International Conference on Intelligent User Interfaces (IUI'02),* ACM Press, pp. 15- 22, ISBN 1-58113-459-2.

Becker, M., W. Drozdzynski, H.U. Krieger, J. Piskorski, U. Schäfer, F. Xu. (2002) „SProUT - Shallow Processing with Typed Feature Structures and Unification” In Proceedings of ICON 2002 - International Conference on NLP, Mumbai, India.

Bering Ch., Drozdzynski W., Erbach G., Guasch C., Homola P., Lehmann S., Li H., Krieger H-U., Piskorski J., Schäfer U., Shimada A., Siegel M., Xu F. and Ziegler-Eisele, D. (2003). „Corpora and Evaluation Tools for Multilingual Named Entity Grammar Development”. In: S.

Neumann and S. Hansen-Schirra (eds.). In: *Proceedings of the Workshop on Multilingual Corpora, Corpus Linguistics Conference*. Lancaster.

Bohnenberger, T., Jameson, A., Krüger, A. and Butz, A. (2002). „User acceptance of a decision-theoretic, location-aware shopping guide". In: Y. Gil and D. Leake (Eds.), *IUI 2002: International Conference on Intelligent User Interfaces* pp. 178-179). New York: ACM.

Bohnenberger, T., Jameson, A., Krüger, A. and Butz, A. (2002). „Location-aware shopping assistance: Evaluation of a decision-theoretic approach". In: *Proceedings of the Fourth International Symposium on Human-Computer Interaction with Mobile Devices,* Pisa, pp. 155-169.

Braun, Christian. 2003. "Parsing German text for syntacto-semantic structures." *Proceedings of the Lorraine*-Saarland *Workshop on Prospects and Advances in the Syntax/Semantics Interface.* Nancy, Frankreich.

Busemann, S., W. Drozdzynski, H-U. Krieger, J. Piskorski, U. Schäfer, H.Uszkoreit, F. Xu (2003). „Integrating Information Extraction and Automatic Hyperlinking". In Proceedings of ACL-2003, 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan

Capstick, J., Declerck, T., Erbach, G., Jameson, A., Jörg, B., Karger, R., Uszkoreit, H., Wahlster, W. and Wegst, T. (2002). „COLLATE: Competence center in speech and language technology". In: *Proceedings of the Third International Conference on Language Resources and Evaluation,* Las Palmas.

Crysmann, B., Frank A., Kiefer B., Müller S., Neumann G., Piskorski J., Schäfer U., Siegel M., Uszkoreit H., Xu F., Becker M. and Krieger H. (2002). „An Integrated Architecture for Shallow and Deep Processing". In: *Proceedings of the 40th Meeting of the Association for Computational Linguistics*. Philadelphia.

Drozdzynski , W., P. Homola, J. Piskorski (2003) "Adapting SProUT to processing Baltic and Slavonic Languages" In Proceedings of IESL'03 Workshop held in conjunction with the RANLP Recent Advances in Natural Language Processing 2003 conference, Borovets, Bulgaria.

Erk, K., Kowalski, A., Padó, S. and Pinkal, M. (2002). "A Corpus Resource for Lexical Semantics". In: *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS)*.

Erk, K., Kowalski, A., Padó, S. and Pinkal, M. (2003a). "A Large Corpus with Extensive Semantic Annotation." Accepted for ACL 2003.

Erk, K., Kowalski, A., Padó, S. and Pinkal, M. (2003b). "Building a Resource for Lexical Semantics." Accepted for the 17th International Conference of Linguists (CIL).

Fliedner, G. (2002). "A System for Checking NP Agreement in German Texts." In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), July 6-12. Companion Volume of the Conference Proceedings. Proceedings of the Student Workshop, Demonstration Abstracts and Tutorial Abstracts.* pp 12-17, University of Pennsylvania, Philadelphia.

Fliedner, Gerhard. 2003. "Tools for building a lexical semantic annotation". Proceedings of the Lorraine-Saarland Workshop on Prospects and Advances in the Syntax/Semantics Interface. Nancy, Frankreich.

Gabsdil, M. (2003). "Clarification in Spoken Dialogue Systems." In: *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*.

Jörg, Brigitte (2003). „Methodologie, Architektur und Umsetzung eines webbasierten Fach-informationssystems mit dem Beispiel Sprachtechnologie“ Magisterarbeit an der Universität des Saarlandes, FR 5.6 Informationswissenschaft.

Karagjosova, E. and Kruijff-Korbayová, I. (2002a). “An analysis of conditional responses in dialogue.” *5th International Conference on TEXT, SPEECH and DIALOGUE (TSD 2002)*.

Karagjosova, E. and Kruijff-Korbayová, I. (2002b). “Conditional responses in information seeking dialogues.” *3rd SIGdial Workshop on Discourse and Dialogue*.

Keller, F., Lapata M. and Ourioupina, O. (2002). “Using the Web to Overcome Data Sparseness.” *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Kray, C., Wasinger, R., Kortuem, G., "Concepts and issues in interfaces for multiple users and multiple devices", *Workshop on Multi-User and Ubiquitous User Interfaces (MU3I) at IUI/CADUI*, 2004.

Krieger , H.-U. (2003) „SDL - A Description Language for Building NLP Systems“ In The Proceedings of HLT-NAACL 2003 Workshop "Software Engineering and Architecture of Language Technology Systems, Edmonton, Canada.

Krieger , H.-U.and J. Piskorski (2003) „Speed-up methods for complex annotated finite state grammars”. DFKI Report.

Krüger, A., Kruppa, M., Müller, C. and Wasinger, R. (2002). “Readapting multimodal presentations to heterogeneous user groups“. In: *Notes of the AAAI-Workshop on Intelligent and Situation-Aware Media and Presentations*, Technical Report WS-02-08, AAAI Press. pp. 46-54.

Krüger, A., Butz, A., Müller, C., Stahl, C., Wasinger, R., Steinberg, K, Dirschl, A., "The Connected User Interface: Realizing a Personal Situated Navigation Service", *Proc. of the International Conference on Intelligent User Interfaces*, 2004.

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R. and Wittig, F. (2001). „Recognizing time pressure and cognitive load on the basis of speech: An experimental study”. In: M. Bauer, P. Gmytrasiewicz and J. Vassileva (Eds.), *UM2001, User Modeling: Proceedings of the Eighth International Conference,* pp. 24-33. Berlin. Springer.

Müller, Ch. (2002). “Multimodal Dialog in a Mobile Pedestrian Navigation System”. In: *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Multi-Modal Dialogue in Mobile Environments*. Kloster Irsee.

Müller Ch. and Wasinger, R. (2002). “Adapting Multimodal Dialog for the Elderly”. In. *Proceedings of the ABIS-Workshop 2002 on Personalization for the Mobile World*. *Hannover. October 9-11*.

Müller, Ch., Wittig, F. and Baus, J. (2003). “Exploiting speech for recognizing elderly users to respond to their special needs”. In: *Proceedings of the Eight European Conference on Speech Communication and Technology (Eurospeech 2003).*

Müller, Ch. and Wittig, F. (2003). “Speech as a Source for Ubiquitous User Modeling”. In: *Proceedings of the Workshop on User Modeling for Ubiquitous Computing in conjunction with User Modeling 2003 Conference*. *Pittsburgh, PA*.

Ourioupina, O. (2002). “Extracting Geographical Knowledge from the Internet.” In: *Proceedings of the ICDM-AM International Workshop on Active Mining 2002.*

Rupp, C.J. (2003). "Prototyping Dialogues for Information Access." Proceedings of the 7th workshop on the Semantics and Pragmatics of Dialogue (DiaBruck 2003). Saarbrücken.

Uryupina, Olga (2003). “Semi-supervised learning of Geographical gazetteers from the Internet.”. *Proceedings of the HLT-NAACL Workshop on the Analysis of Geographic References*. Edmonton, Kanada.

Hans Uszkoreit, Brigitte Jörg, Gregor Erbach (2003) "An Ontology-based Knowledge Portal for Language Technology" In: *Proceedings of ENABLER/ELSNET Workshop "International Roadmap for Language Resources"*, Paris.

Wahlster, W. (2002). „Disambiguierung durch Wissensfusion: Grundprinzipien der Sprachtechnologie“. In: *KI - Künstliche Intelligenz, Heft 1*.

Wahlster, W. (2003): “Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture, and Facial Expression”. In: Günter, A., Kruse, R., Neumann, B. (eds.): KI 2003: Advances in Artificial Intelligence. Proceedings of the 26th German Conference on Artificial Intelligence, September 2003, Hamburg, Germany, Pages 1 - 18, Berlin, Heidelberg: Springer, LNAI 2821, 2003.

Wasinger, R., Stahl, C. and Krüger, A. (2003). “Robust speech interaction in a mobile environment through the use of multiple and different media input types“. In: *Proceedings of EuroSpeech*.

Wasinger, R., Stahl, C. and Krüger, A. (2003). “M3I in a Pedestrian Navigation & Exploration System“. In: *Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices*.

Wasinger, R., Kray, C. and Endres, C. (2003). “Controlling multiple devices“. In: *Physical Interaction (PI03) Workshop on Real World User Interfaces*.

Wasinger, R., Oliver, D., Heckmann, D., Braun, B., Brandherm, B., Stahl, C., "Adapting Spoken and Visual Output for a Pedestrian Navigation System, based on given Situational Statements", *ABIS Workshop on adaptivity and user modelling in interactive software systems*, 2003.

Wirschum, N. (2003). „Erleichterung der Informationssuche im Internet“. Master's thesis, Saarland University.

Yao, T., Ding, W. and Erbach, G. (2002). “Correcting word segmentation and part-of-speech tagging errors for Chinese named entity recognition”. In: Günter Hommel and Shen Huanye (Eds.): *The Internet Challenge: Technology and Applications.* pp. 29-36, Kluwer Academic Publishers, Dordrecht.

Yao, T., Ding, W. and Erbach, G. (2002). “Repairing Errors for Chinese Word Segmentation and Part-of-Speech Tagging”. In: *Proc. of the First International Conference on Machine Learning and Cybernetics 2002 (ICMLC 2002).* pp. 1881-1886. Beijing.

Yao, T., Ding, W. and Erbach, G. (2003). “CHINERS: A Chinese Named Entity Recognition System for the Sports Domain”. In: *Proc. of the Second SIGHAN Workshop on Chinese Language Processing (ACL 2003 Workshop).* Sapporo.

Zheng, Z. (2002). “Rule-based Sentence Segmentation for HTML/TEXT Documents”. *The Thirteenth meeting of Computational Linguistics in the Netherlands (CLIN 2002).* Groningen.

Zheng, Z. (2002). „Natural Stemming Derived from Porter's Algorithm”. *The Thirteenth meeting of Computational Linguistics in the Netherlands (CLIN 2002)*. Groningen.

Zheng, Z. and Erbach, G. (2002). „Specialized search in linguistics and languages”. *XI International Conference on Computing (CIC 2002). Mexico City.*

Zheng, Z., Huang, H. and Schmeier, S. (2002). „Deploying Web-based Question Answering System to Local Archive”. *Fifth International Conference on TEXT, SPEECH and DIALOGUE (TSD 2002). Brno.*

Zheng, Z. and Erbach, G. (2003). „Using Specialized Knowledge in Automated Web Document Summarization”. *The Fifth International Conference on Enterprise Information Systems (ICEIS 2003). Angers.*

Zheng, Z. (2003). “Answer Bus News Engine”. *The Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003). Budapest. April 12-17*.

Vorträge

Erbach, G.: Special Topic Session "Collocations and Information Management", Workshop on Computational Approaches to Collocations, Universität Wien, 23. 07. 2002 (invited speaker).

Erbach, G.: “Collate: Competence center in speech and language technology”. Third International Conference on Language Resources and Evaluation. 30.05.2002.

Erbach, G.: “Web- und Videokonferenzen als Arbeitsmittel und Forschungsgegenstand”. Fachtagung Web und Videoconferencing im Universitätsumfeld. Saarbrücken, 04.06.2003.

Gregor Erbach "An Ontology-based Knowledge Portal for Language Technology": ENABLER/ELSNET Workshop "International Roadmap for Language Resources", Paris, 28.08.2003

Jameson, A.: „Wie wähle ich eine Telefonnummer, während ich die Treppe 'runterlaufe?“ Antrittsvorlesung. Philophische Fakultät, Universität des Saarlandes, 11.11.2002.

Pinkal, M.: “Semantics, Pragmatics, and Dialogue Applications”. (Invited speaker). EDILOG 2002, 6th Workshop on the Semantics and Pragmatics of Dialogue, The University of Edinburgh. 4.-6.09.2002.

Pinkal, Manfred. „Using a Large-Scale Lexical Semantics Resource for Information Acquisition“. Vortrag am ISI, Marina del Rey, CA.

Pinkal, Manfred. „A Frame-based Large-Scale Lexical Semantics Resource for German“. Vortrag am ICSI, Berkeley, CA.

Pinkal, Manfred. „Supporting Lexical Semantic Acquisition by Deep and Flat Syntactic Information“. Palo Alto Research Center (PARC), Palo Alto, CA.

Schofield, E. J. and Zheng, Z.: “A Speech interface for open-domain question-answering”. 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003). Sapporo, Japan. 7-12.07.2003.

Uszkoreit, H.: "Sprachtechnologie für das Wissensmanagement", Linguistisches Kolloquium, Universität München. 30.01.2002.

Uszkoreit, H. "Ein integrierter Ansatz zur datenorientierten Sprachforschung und -technologie", Kolloquium Korpusannotierung/Linguistische Datenbanken, Institut für Dolmetschen und Übersetzen, Universität des Saarlandes. 15.02.2002.

Uszkoreit, H. "Können Maschinen verstehen, was Kunden wünschen? Sprachtechnologie im Call Center", CallCenterWorld 2002, Berlin. 28.03.2002.

Uszkoreit, H.: "Language Technology for Knowledge Management", Euroscript Workshop "Megatrends in Language-Competent Web-Services", Luxembourg. 05.06.2002.

Uszkoreit, H.: Keynote Lecture "Language Technologies for Knowledge Applications", PorTAL 2002, Faro Portugal. 23.06.2002.

Uszkoreit, H.: "Future Directions in Interactivity", Workshop "Future Directions in Interactivity, Content and Knowledge" European Commission, Luxembourg. 27.06.2002.

Uszkoreit, H.: Keynote Lecture "New Chances for Deep Linguistic Processing", COLING 2002. Taipei. 26.08.2002.

Uszkoreit, H.: "The Semantic Web – A Challenge for Language Technology" Panel Presentation and Moderation at COLING 2002. Taipei. 30.08.2002.

Uszkoreit, H.: "Area Report Information Retrieval and Information Extraction", Workshop "Towards a Roadmap for Computational Linguistics" at COLING 2002. Taipei. 31.08.2002.

Uszkoreit, H.: "The ELSNET Roadmap Initiative", Workshop "Towards a Roadmap for Computational Linguistics" at COLING 2002. Taipei. 31.08.2002.

Uszkoreit, H.: Keynote Lecture "Combining Deep and Shallow Processing", International Workshop "Treebanks and Linguistic Theories". Sozopol. 20.09.2002.

Uszkoreit, H.: "The Virtual Information Center LT World", NorDokNet Meeting, Göteborg. 21.10.2002.

Uszkoreit, H.: "Impossibilities Creatively Combined - New Forms of Scientific Collaboration", Norwegian National Language Technology Conference. Bergen. 25. 10.2002.

Uszkoreit, H.: "Embedded Language Technology", Workshop "Beyond Language Technology" at the IST 2002 Conference, Copenhagen. 05.11.2002.

Uszkoreit, H.: "The Evolution of Language Technology", LISA Forum Europe 2002 (Localization Industry Standards Association), Heidelberg. 06.11.2002.

Uszkoreit, H.: “Shallow and Deep Processing.” Center for Sprogteknologi, Copenhagen. 27.11.2002.

Uszkoreit, H.: Keynote Lecture "Shallow and Deep Processing in Language Technology", Aslib Conference "Translating and the Computer 24", London. 21.11.2002.

Uszkoreit, H.:"European Web of Culture, History and Science" (with Jürgen Renn), Workshop "Mapping the Future", Organized by the Unit for Preservation and Enhancement of Cultural Heritage, Luxembourg, 28 Januar 2003

Uszkoreit, H.:"Hybride Methoden in der Computerlinguistik", Wissenschaftliches Symposion Computerlinguistik, Bonn, 15 January 2003

Uszkoreit, H.: "Research and Services of the German Competence Center for Language Technology", International Symposium "Ubiquitous Networked Media Computing", Nara Institute of Science and Technology, Nara, 18.03.2003.

Uszkoreit, H.:Keynote Lecture"Hybrid Methods for Language Processing", Annual Conference of the Association for Natural Language Processing of Japan, Yokohama, 20 March 2003

Uszkoreit, H.:"A Transatlantic Cooperation in Crosslingual Information Retrieval" International Symposion on Medical Informatics "Pioneers and MuchMore", University of Frankfurt, 16.05.2003

Uszkoreit, H.:"The Role of Linguistics for the Future of Natural Language Processing" Workshop "Leitlinien für die Heidelberger Computerlinguistik", Universität Heidelberg, 20 June 2003

Uszkoreit, H.:"Digitale Gedächtnisse: Konzepte, Methoden, Anwendungen", Berliner Bibliothekswissenschaftliches Kolloquium, Humboldt Universität, Berlin, 15 July 2003

Uszkoreit, H.:Keynote Lecture "Language Technology for Digital Memories", Conference "Recent Advances in Natural Language Processing", RANLP 2003, Borovets, 11 September 2003

Uszkoreit, H.:"Hyperlinking for Weaving Digital Cultural Memories", Lund Technology Days of the International, Initiative European Cultural Heritage Online, Lund, 18 September 2003

Uszkoreit, H.:"From Lab to Market", Workshop on Language and Communication Technology, Trento, 17.10.2003

Uszkoreit, H.:"Education in a Transdiscipline", Workshop on Language and Communication Technology, Trento • 17 October 2003

Uszkoreit, H.: "Language Technology for Associative Memories and the Semantic Web", LangTech 2003, Paris, 24.11.2003

Wasinger, R., C. Stahl and A. Krüger: "*M3I in a Pedestrian Navigation & Exploration System*" In: Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices. 2003.

Wahlster, W.: "Personalized Web Interaction". IIP-2002. IFIP World Computer Congress 17th Edition. Montreal. 29.08.2002.

Wahlster, W.: "Language Technologies for the Mobile Internet Era". LangTech 2002, Berlin. 27.09.2002.

Wahlster, W.: "Mobile Multimodal Dialogue Systems". Interact 2003, Zürich, 30.09.2003

Wahlster, W.: "Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture and Facial Expression". 26th Annual German Conference on Artificial Intelligence 2003 (KI 2003), Hamburg, 16.09.2003

Yao, T.: "Correcting word segmentation and part-of-speech tagging errors for Chinese named entity recognition". TU-Berlin Workshop „The Internet Challenge: Technology and Applications", Berlin. 08.10.2002.

Yao, T.: "Repairing Errors for Chinese Word Segmentation and Part-of-Speech Tagging". First International Conference on Machine Learning and Cybernetics 2002 (ICMLC 2002). Beijing. 04.11.2002.

Yao, T.: "Chinese Named Entity Recognition on Sports Domain: Techniques and Implementation". Shanghai Jiao Tong University. Shanghai. 13.11.2002.

T. Yao. CHINERS: A Chinese Named Entity Recognition System for the Sports Domain. Second SIGHAN Workshop on Chinese Language Processing (ACL 2003 Workshop). Sapporo, Japan. July 12, 2003.

Zheng, Z., Huang, H. and Schmeier, S.: Deploying Web-based Question Answering System to Local Archive. Fifth International Conference on TEXT, SPEECH and DIALOGUE (TSD 2002). Brno. 9-12.09.2002.

Zheng, Z.: "Rule-based Sentence Segmentation for HTML/TEXT Documents". The Thirteenth meeting of Computational Linguistics in the Netherlands (CLIN 2002). Groningen. 29.11.2002.

Zheng, Z.: "Natural Stemming Derived from Porter's Algorithm". The Thirteenth meeting of Computational Linguistics in the Netherlands (CLIN 2002). Groningen. 29.11.2002.

Zheng, Z.: "Specialized search in linguistics and languages". XI International Conference on Computing (CIC 2002). Mexico City. 25-29.11.2002.

Zheng, Z.: "Question Answering Using Web News as Knowledge Base". The Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003). Budapest. 12-17.04.2003.

Zheng, Z.: "Using Specialized Knowledge in Automated Web Document Summarization". *T*he Fifth International Conference on Enterprise Information Systems (ICEIS 2003). Angers. 23-26.04. 2003.

Zheng, Z.: "AnswerBus News Engine". The Twelfth International World Wide Web Conference (WWW 2003). Budapest. 20-24.05.2003.

Zheng, Z.: "Design, Development and Implementation of Question answering Systems" (Tutorial). Third International Conference on Intelligent Systems Design and Applications (ISDA 2003). Tulsa. 10-13.08.2003.