Robustes Verstehen gesprochener Sprache in einem multimodalen Roboter-Szenario

Sonja Hüwel

Dipl.-Inform. Sonja Hüwel
AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld
email: shuewel@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung
des akademischen Grades Doktor-Ingenieurin (Dr.-Ing.).
Der Technischen Fakultät der Universität Bielefeld
am 22.01.2006 vorgelegt von Sonja Hüwel,

Gutachter:

Dr. Britta Wrede, Universität Bielefeld
Prof. Dr. Henning Lobin, Justus-Liebig Universität Gießen

Prüfungsausschuss:

Prof. Dr. Ipke Wachsmuth, Universität Bielefeld
Dr. Britta Wrede, Universität Bielefeld
Prof. Dr. Henning Lobin, Justus-Liebig Universität Gießen
Prof. Dr. Franz Kummert, Universität Bielefeld
Dr. Thomas Hermann, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier nach ISO 9706

Robustes Verstehen gesprochener Sprache in einem multimodalen Roboter-Szenario

Der Technischen Fakultät der Universität Bielefeld

zur Erlangung des Grades

Doktor-Ingenieurin

vorgelegt von

Sonja Hüwel

Bielefeld – Dezember 2006

Danksagung

Als erstes bedanke ich mich bei meiner Gutachterin Dr. Britta Wrede für ihre Unterstützung. Ihr Interesse an meiner Arbeit und die zahlreichen Diskussionen waren überaus hilfreich. Ebenso bedanke ich mich bei meinem Zweitgutachter Prof. Dr. Henning Lobin. Er war mir gerade in der schwierigen Anfangsphase eine sehr hilfreiche Stütze.

Als nächstes möchte ich mich bei Prof. Dr. Gerhard Sagerer bedanken, der mich überhaupt zur Promotion ermutigte. Er hat innerhalb seiner Arbeitsgruppe einen offenen Raum geschaffen, in dem auch sehr interdisziplinäre Arbeiten realisierbar sind. Ohne diese vielfältigen Austauschmöglichkeiten, wäre diese Arbeit nicht enstanden. Ein besonderer Dank geht an meine Bürokollegen sowie an die übrigen Mitglieder der AG „Angewandte Informatik“. Sie trugen dazu bei, eine sehr angenehme Arbeitsatmosphäre zu schaffen und standen mir für fachliche Diskussionen und technische Unterstützung immer zur Seite. Unter anderem hätten ohne sie die Experimente mit dem Roboter BIRON gar nicht stattfinden können.

Besonders bedanken möchte ich mich bei meiner Familie, die mich in vielen Dingen entlastet und mir den „Rücken frei gehalten“ hat für meine Arbeit. Ich weiß, dass sie im Notfall immer für mich da ist und das gibt mir die Kraft, auch schwierige Situationen zu bewältigen. Bei Cornelia möchte ich mich für den „warmen Platz am Kamin“ bedanken. Sie hat mich aus meinem Trott herausgeholt und mir die Augen offengehalten für die anderen wichtigen Dinge im Leben. Sie gab mir, wie auch meine anderen Freunde, wichtigen mentalen Rückhalt. Mein größter Dank gilt Björn. Er war mein größter Kritiker und zugleich mein größter Fan. Mit ihm habe ich viele intensive Diskussionen geführt und dadurch neue Impulse bekommen. Er hat mich in der gesamten Zeit begleitet.

Inhaltsverzeichnis

1. Einleitung

In den letzten Jahren mehren sich in den Medien Informationen über Robotersysteme, die dem Menschen zur Seite stehen und ihn im Alltag unterstützen sollen. In diesem Zusammenhang keimen vielfältige Ideen auf, in welchen Bereichen diese Roboter zum Einsatz kommen können. Diese Ideen inspirieren Wissenschaft und Forschung, wobei besonders Einzelaspekte große Beachtung finden. Der Einsatzbereich von Roboterassistenten oder auch *Robot Companions* ist groß: Sie werden für den Bereich der Assistenz in Krankenhäusern oder Altersheimen entwickelt [Spi01, Cap92, Mon02, Par01], sollen im Haushalt helfen [Sch99, Gra04, Rog02], als Boten dienlich sein [Kri01, Tsc01, Böh98] oder auch als „mobiler“ Informationsbringer [Nou99, Thr00, Tom02, Krö03] fungieren. Daneben hat die Spielzeugindustrie immer mehr Interesse an der Entwicklung von Spielzeug-Robotern, wie z. B. Sonys AIBO®, Hasbros My-Real-Baby® oder NECs PaPeRo® [Bar01]. Da sie zumeist interaktive Fähigkeiten besitzen und für sie das Wahrnehmen und Äußern von Emotionen sehr wichtig ist, werden diese spielzeugähnlichen Systeme auch für therapeutische Maßnahmen [Shi03] eingesetzt.

Bislang jedoch handelt es sich bei der Mehrzahl der sich im Einsatz befindlichen Roboter um Industrieroboter [Gug99], auch wenn der Trend deutlich in Richtung Service-Robotik und persönliche Robotik zeigt. Die Industrieroboter sind mit ihren Fähigkeiten genau auf ihre jeweilige Spezialaufgabe ausgerichtet, besitzen einen festen Bewegungsablauf und nur wenige Sensoren. Sie sind nicht dafür ausgelegt, mit einem Menschen zu interagieren oder ihn in alltäglichen Aufgaben zu unterstützen.

Um die Entwicklung der Roboter in Richtung kommunikative und persönliche Assistenz voranzutreiben, ist die Beteiligung vieler Forschungsbereiche notwendig. Die Entwicklung technischer Systeme für den Einsatz in der Mensch-Maschine-Kommunikation bedarf nicht nur Kenntnisse über die technische Realisierung des Systems. Vielmehr findet ein komplexes Zusammenspiel zwischen Wissen aus sehr unterschiedlichen Disziplinen wie der Informatik, Elektrotechnik, Design, Linguistik, Soziologie u. a. statt. Oftmals werden in der aktuellen Forschung jeweils nur Einzelaspekte betrachtet, jedoch nicht der Gesamtzusammenhang vieler Gesichtspunkte berücksichtigt. Dabei kann schon das Fehlen eines wichtigen Bereiches das Gesamtkonzept stark einschränken.

Gerade bei der Entwicklung mobiler Roboter mit sozialen Fähigkeiten wird deutlich, wie wichtig es ist, alle Aspekte in der Konzeption mit einzubeziehen. Dafür ist eine andere Denkweise notwendig: das Verständnis für die Probleme der anderen Disziplinen und die Entwicklung gemeinsamer Lösungen unter interdisziplinären Gesichtspunkten. Die hier vorgestellte Arbeit gibt einen Einblick in die Komplexität und Vielfältigkeit bei der Entwicklung eines interaktiven Systems,

genauer gesagt eines *Robot Companions*. Der Fokus liegt dabei jedoch auf der Entwicklung einer Sprachverstehenskomponente, bei der sowohl die technischen Zusammenhänge, in die sie eingebunden ist, als auch die sozialen und kommunikativen Rahmenbedingungen berücksichtigt wurden.

1.1. Motivation

Gegenwärtig ist es meist noch so, dass der Mensch sich den Gegebenheiten seiner Umwelt anpasst. Das bedeutet auch, dass er lernen muss, mit der vorhandenen Technik umzugehen. Jedoch sind viele Menschen überfordert, neue Denkweisen anzunehmen und sich an die vielen Systeme mit den doch sehr vielfältigen Funktionalitäten und dem unterschiedlichen Aussehen zu gewöhnen. Da kann schon ein Videorekorder oder ein Fahrkartenautomat ein unüberwindliches Hindernis sein. Werden neue Techniken eingeführt, muss man immer berücksichtigen, welche Menschen mit ihnen interagieren werden. Gerade bei der Entwicklung von sozialen Robotersystemen wird dies offensichtlich: Die Aufgaben, die diese Systeme übernehmen können, sind vielfältig. Sie können für verschiedenste Aufgaben und daher auch für sehr unterschiedliche Benutzergruppen zum Einsatz kommen. Insbesondere bei der Unterstützung älterer Menschen oder Menschen mit Behinderungen scheinen mobile Robotersysteme eine Möglichkeit zu bieten, selbständiger und freier leben zu können [Gra04]. Auch in Museen [Bur98, Sch01] oder anderen Gebäuden wie Bürogebäuden [Aso01] oder Baumärkten [Böh03], in denen viele Menschen zusammenkommen, kann ein interaktiver Roboter sinnvoll eingesetzt werden.

Man kann nicht erwarten, dass diese Menschen eine eigene für das System entwickelte Kommandosprache lernen oder sogar die Eingabe der Kommandos per Tastatur vornehmen. Daher ist es ein Anliegen in der Forschung, die Kommunikation so intuitiv und natürlich wie möglich zu gestalten. Das System soll sich mit seinen Fähigkeiten dem Menschen anpassen und nicht der Mensch an das System.

Offensichtlich ist es für die gesamte Interaktion hilfreich, wenn die Benutzer eine positive Einstellung zum Robotersystem entwickeln, ihn bei sich zuhause dulden und ihn vor allem als Kommunikationspartner akzeptieren können. Folglich sind eine Menge von entsprechenden Verhaltensweisen in das Gesamtsystem zu integrieren und vor allem die interaktiven Fähigkeiten des Roboters möglichst intuitiv und natürlich zu realisieren.

Das bedeutet für die Gestaltung der Interaktion, ein breites Spektrum kommunikativer Mittel zu erlauben: den Einsatz von Spontansprache, von Gestik und Mimik sowie das zusätzliche Nutzen von Informationen aus der Umgebung. Es muss auch damit gerechnet werden, dass verschiedene Personen – aufgrund ihrer sehr individuellen Vorerfahrungen mit technischen Systemen – auf unterschiedliche Weise mit dem System interagieren und auch auf unterschiedliche Sprachstile zurückgreifen.

Für die sprachliche Interaktion mit einem künstlichen System werden verschiedene Komponenten benötigt: Die Spracherkennung erhält das akustische Signal und transformiert es in eine

symbolische Form (Wortkette). Danach muss die Wortkette in eine semantische Form umgewandelt werden, die dann von einem Dialogmanager interpretiert werden kann. Dazu wird eine Sprachverstehenskomponente benötigt, die unter Verwendung linguistischer Wissensbasen (z. B. Lexikon, syntaktisches Wissen in Form einer Grammatik und semantisches Wissen über Wortbeziehungen) aus der Wortkette eine kohärente semantische Struktur erzeugt. Auf Basis der semantischen Informationen kann das Dialogsystem dann entsprechend reagieren (z. B. auf eine Anfrage antworten oder eine Handlung ausführen).

1.2. Zielsetzung

Das Ziel der hier beschriebenen Arbeit ist die Entwicklung einer Sprachverstehenskomponente für den Einsatz in Robotersystemen mit interaktiven und sozialen Fähigkeiten. Dieses System soll die Kommunikationsfähigkeit des Robotersystems fördern, die Funktionalität anderer Systemkomponenten soweit wie möglich unterstützen und in die bestehende Architektur integriert werden. In dieser Arbeit werden die vielfältigen Anforderungen, vor die ein komplexes sprachverstehendes System in diesem Kontext gestellt ist, angesprochen und berücksichtigt.

Der zentrale Gedanke bei der Entwicklung der Komponente ist, eine möglichst intuitive Kommunikation mit dem Roboter zu ermöglichen und zu fördern. Daher stellt sich zunächst die Frage, was unter „natürlicher Kommunikation“ zwischen Mensch und *Robot Companion* zu verstehen ist. Als Annahme liegt dabei zugrunde, dass Menschen mit Robotern aufgrund fehlender Vorerfahrungen mit einem Robotersystem zunächst einmal ähnlich kommunizieren wie mit einem anderen Menschen [Nas00]. Eine alternative Annahme bei der Kommunikation mit technischen Systemen ist die Verwendung des so genannten „Computer-Talks“, d. h. die Benutzter reduzieren ihre Sprache auf notwendige Informationen [Kra92]. Aufgrund mangelnder Studien können bislang keine klaren Aussagen über die Art der Interaktion mit sozialen Robotern getroffen werden. Experimente und Untersuchungen von Mensch-Roboter-Interaktionen sind daher ein wichtiger Bestandteil für die Entwicklung eines intuitiven Robotersystems, vor allem für die Aspekte der Kommunikation. Nur so lassen sich die Eigenheiten der Mensch-Roboter-Dialoge bestimmen und somit die Rahmenbedingungen für die Verarbeitung der Äußerungen in diesem Kontext festlegen.

Daran anknüpfend kann dann das benötigte Weltwissen und der Sprachumfang für ein konkretes Szenario (vgl. Kap. 6) bestimmt werden, auf das das automatische Sprachverstehen aufbaut. Mehr noch, der gesamte Verarbeitungsprozess muss im Hinblick auf diese Besonderheiten und Anforderungen sowohl aus der kommunikativen Sicht als auch aus der technischen Sicht abgestimmt sein.

Letztendlich soll ein funktionierendes Robotersystem entstehen, in dem die verschiedenen Einzelkomponenten effizient zusammenarbeiten. Die einzelnen Komponenten sollen sich gegenseitig unterstützen, Schwachstellen berücksichtigen und sie in ihre Verarbeitungskonzepte einbeziehen. Der Gewinn des Gesamtsystems ist demnach mehr als die „Summe seiner Teile“. Bei der

Entwicklung der Sprachverstehenskomponente des Roboters fließen daher nicht nur das Wissen über die Domäne, sondern ebenfalls die Bedingungen für die direkten Schnittstellen sowie das Wissen über den gesamten Informationsfluss ein. Unter Berücksichtigung all dieser Informationen wird in dieser Arbeit eine Sprachverstehenskomponente entwickelt mit dem Ziel, eine dem Menschen möglichst angenehme und hilfreiche Kommunikation zu ermöglichen und ihn letztendlich dadurch in seinen Aufgaben zu unterstützen.

1.3. Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich konzeptionell in drei Teile. Teil I beschreibt die theoretischen Grundlagen der Arbeit und gibt einen Überblick über den aktuellen Forschungsstand und die bereits vorhandenen Systeme im Bereich Sprachverstehen für mobile Robotersysteme. Darin werden zunächst in Kapitel 2 die Grundlagen der Mensch-Roboter-Kommunikation beschrieben. In Kapitel 3 werden Ansätze zur Wissensrepräsentation sowie Sprachverarbeitungsmechanismen vorgestellt. Anschließend werden in Kapitel 4 ausgewählte Dialogsysteme unter Einbeziehung der Aspekte Situiertheit, Spontansprache und Einsatzfähigkeit in mobile Robotersysteme diskutiert. Ebenfalls wird der Stand der Forschung im Bereich Sprachverstehen für mobile *Robot Companions* beschrieben.

Teil II erläutert die besonderen Rahmenbedingungen kommunikativer Roboter, die das sprachverstehende System beachten muss. Zum einen sind das die Besonderheiten eines mobilen Robotersystems, die beim Entwurf der Verstehenskomponente beachtet werden müssen sowie die Anforderungen an die Schnittstellen zu anderen Komponenten. Diese Aspekte werden speziell für den Roboter BIRON [Haa04] – den *Bielefeld Robot Companion* – in Kapitel 5 beschrieben. Zum anderen besitzt gerade das Korpus der Mensch-Roboter-Kommunikation besondere Merkmale, die in Kapitel 6 diskutiert werden. Aus diesen Rahmenbedingungen ergeben sich Anforderungen und Designkriterien, die in Kapitel 7 erläutert werden.

In Teil III wird das entwickelte Konzept für das Verstehen von Sprache für *Robot Companions* vorgestellt. Zunächst wird das Konzept der Wissensrepräsentationen, d. h. das Lexikon und die *Situierten Semantischen Einheiten*, in Kapitel 8 beschrieben, die insbesondere für die Repräsentation situierter Spontansprache geeignet sind. Anschließend wird in Kapitel 9 der dazu passende Verarbeitungsmechanismus dargestellt und die Anbindung an die Roboterplattform BIRON erläutert. In Kapitel 10 wird die Evaluation des Gesamtkonzepts dargelegt, die die Verbindung der Wissensbasen mit dem Verarbeitungsmechanismus bewertet und darüber hinaus die Praktikabilität des Konzeptes für *Robot Companions* am Beispiel des Roboters BIRON belegt.

Die vorliegende Arbeit schließt mit einer Diskussion des Konzeptes in Kapitel 11 und einer Zusammenfassung in Kapitel 12 ab.

2. Theoretische Grundlagen der Mensch-Roboter-Kommunikation

In diesem Kapitel werden Theorien und Begriffe aus der Kommunikationstheorie vorgestellt, die für die Interaktion zwischen einem Roboter und einem Menschen von Bedeutung sind. Sie sind notwendig, um die Mechanismen und Wechselwirkungen in Kommunikation und Interaktion besser zu verstehen. Insbesondere werden die Verbindungen zwischen der Interaktion der beteiligten Personen, dem Wissen über kommunikatives Verhalten und der Umgebung betrachtet. Diese Überlegungen und Modelle flossen in den Entwicklungsprozess der kommunikativen Systemkomponenten und insbesondere der Sprachverstehenskomponente für den in dieser Arbeit beschriebenen *Robot Companion* BIRON [Haa04] mit ein.

2.1. Allgemeine Begriffsdefinitionen

Kommunikation ist eine Form sozialer Interaktion und Informationsübertragung [Dor87]. Der Prozess der Informationsübertragung besteht aus den Komponenten *Kommunikator*, das ist der Sender der Information, und *Kommunikant*, das ist der Empfänger der Information, den *Kommunikationsmitteln* sprachlicher oder nicht-sprachlicher Art und den *Kommunikationsinhalten* ([Dor87] S. 343). *Kommunikation* bezeichnet ganz allgemein Vorgänge, in denen die Information einer Person als auf einen bestimmten Empfänger bezogen betrachtet wird. Das Kontextwissen über die Personengruppe, an die sich die Informationen richten, fließt demnach immer in den Prozess mit ein.

Nach Watzlawick [Wat69] ist Kommunikation ein allgemeiner sozialer Prozess. Dabei wird die Wirkung auf die *Beziehung* zwischen den Kommunikationspartnern im Verlauf eines Kommunikationsprozesses betrachtet. Sowohl die sprachlichen Bestandteile als auch die non-verbalen Begleiterscheinungen schließt Watzlawick in den Kommunikationsbegriff mit ein. Kommunikation ist ein Rückkopplungssystem, bei der die Äußerung des einen Partners auf den anderen wirkt und dessen Reaktionen wiederum auf den ersten zurückwirken. Das Rückkopplungssystem schließt auch Fehlinterpretationen im Modell mit ein.

Handlung ist eine oft komplexe Abfolge von koordinierten und umweltbezogenen Bewegungen, die ein Individuum ausführt ([Dor87] S. 270). Die Handlung hebt sich vom reinen Verhalten dadurch ab, dass sie auf das Erreichen eines Zieles gerichtet ist, sie ist also *intentional*. Da die Intention eines Individuums nicht von außen erkennbar ist, kann nur vermutet werden, warum sich der andere so verhält. Beobachtbar ist nur das *Verhalten*, in dem sich die Handlung äußert.

Kommunikation ist wie Handlung intentional und orientiert sich an der Umwelt, sie kann daher auch als *Sprachhandeln* verstanden werden (siehe auch Abschnitt 2.5). Es können gemäß der Definition auch Handlungen als Kommunikation fungieren, z. B. das Zeigen auf einen Gegenstand oder das Zucken mit der Schulter.

Kooperation ist nach Dohmen die Zusammenarbeit verschiedener Individuen, die auf der Basis gemeinsamer Regeln (für Kommunikation, Zuständigkeiten usw.) auf ein gemeinsames Ziel hinarbeiten [Doh94]. Kooperation ist also die Realisierung einer Tätigkeit durch gemeinsames Handeln. Es beruht demnach auf dem Austausch von Informationen als auch auf Handlungsaktionen. Kommunikation kann zur Unterstützung von Kooperation eingesetzt werden, z. B. in Form einer gemeinsamen Absprache.

Im **Mensch-Roboter-Kontext** hat Kommunikation vor allem zwei Aufgaben. Einerseits vermittelt es die Intention einer Handlung. Andererseits dient es dem Austausch und der Angleichung der Erfahrungswelten der Interaktionspartner. Dafür ist es wichtig, dass der Sprecher und der Hörer die Möglichkeit haben, die innere Erfahrungswelt zu überprüfen, zu adaptieren oder Missverständnisse mitzuteilen. Zusätzlich wirken die Annahmen über die Fähigkeiten des Roboters und das Erscheinungsbild auf die Kommunikationsgestaltung aus. Der Interaktionspartner entwickelt eine Beziehung zum Roboter aufgrund von inneren Bewertungen, des Aussehens und der Reaktionen des Roboters, was das kommunikative Verhalten stark prägt. Ein Roboter kann noch so ausgefeilte kommunikative Fähigkeiten besitzen – er wird dennoch eine gestörte Kommunikation hervorrufen, wenn er durch sein Äußeres den Eindruck erweckt, „dumm“ zu sein, oder der Interaktionspartner vor ihm Angst hat. Bisher ist allerdings nur wenig darüber bekannt, was Personen wirklich über ihre Roboter-Interaktionspartner annehmen und welche Faktoren einen Einfluss haben [Ten03]. In Abschnitt 2.2 wird deshalb näher auf die Besonderheiten der Kommunikation mit einem künstlichen Interaktionspartner eingegangen.

Als **natürlich und intuitiv** werden häufig schon Systeme mit grafischem Interface bezeichnet (wie z. B. in [Koi00]) . Das zeigt deutlich, wie ungenau dieser Begriff ist. Ob ein System intuitiv ist, hängt jedoch nicht alleine von der Ausgestaltung eines Systems ab, sondern auch von den Rahmenbedingungen, innerhalb derer es eingesetzt wird (von der Benutzergruppe, der Umgebung, usw.). In der hier vorgestellten Arbeit wird der Begriff „natürlich“ weitestgehend vermieden. Denn es ist nicht eindeutig definierbar, was genau es bedeuten soll, wenn ein künstliches System natürliche Interaktionsfähigkeiten ermöglicht. Genauso bleibt auch die Frage nicht aus, ob der Mensch nicht ohnehin schon immer natürlich agiert: Wie muss eine Handlung einer Person beschaffen sein, um sie *nicht natürlich* zu nennen? Dennoch sind die Begriffe „natürlich“ und „intuitiv“ wichtig, um das Ziel zu verdeutlichen, eine möglichst reibungslose und angenehme Interaktion zwischen Mensch und Robotersystem zu ermöglichen.

2.2. Aspekte der Mensch-Roboter-Kommunikation

Stehen sich Menschen gegenüber, machen sie bestimmte Annahmen über das Verhalten der anderen, über deren Sicht auf die Welt und deren kognitive Fähigkeiten. Danach richten sie ihr (Sprach-)Handeln aus. Redet ein Erwachsener mit einem Kind, so ändert sich beispielsweise sein Sprachgebrauch, er verwendet einfachere Wörter und eine weniger komplexe Grammatik. Steht ein Mensch einer älteren Person gegenüber, spricht er häufig lauter und langsamer, damit sie ihn besser verstehen kann. Ebenso werden Mimik und Gestik deutlicher.

Doch nicht nur das Gegenüber bestimmt die Art zu handeln. Ein weiterer Aspekt, der in das Verhalten von Menschen einfließt, ist ihre Umgebung. Das Verhalten wird wesentlich von den besonderen Eigenschaften der jeweiligen Umgebung bestimmt. Beispielsweise unterscheidet sich der Sprachstil beim Telefonieren wesentlich von der Kommunikation von Angesicht zu Angesicht. Beim Telefonieren fließen allein die sprachlichen Informationen im Gespräch mit ein, u. a. wird die Mimik des Gegenübers nicht gesehen. Die Telefonierenden passen sich daher bei der Kommunikation an die spezielle Situation an.

Die Unterschiede zwischen verschiedenen Gesprächssituationen wurden in den Studien von Barker [Bar68] in der Behavior-Setting-Theorie festgehalten. Zentraler Aspekt dieser Theorie ist die Betrachtung der Wirkung von Umgebungseigenschaften auf das Verhalten von Personen. Bei Feldstudien zeigte sich, dass die Beschaffenheit der Orte einen entscheidenden Einfluss auf das Verhalten von Individuen hat. Umgebungsvariablen, die in Mensch-Roboter-Szenarien eine Rolle spielen, sind u. a. physikalische Objekte, Zeitdauer, funktionale Rolle der Beteiligten und Motivation. Beispielsweise kann ein Objekt in einer bestimmten Umgebung spezielles Handlungswissen vermitteln. Dafür werden in der Behavior-Setting-Theorie so genannte standardisierte Verhaltensmuster, sogenannte „action patterns“ (Handlungsmuster), definiert.

In alltäglichen Kommunikationssituationen kann Wissen über bereits bekannte Situationen genutzt werden, was sich in Form von Annahmen aus den bisherigen Erfahrungen manifestiert. Was passiert jedoch, wenn Personen zum ersten Mal einem realen Roboter gegenüber stehen? Dann können sie ihr Verhalten nicht auf vorherige Erfahrungen mit Robotern aufbauen. Ihre Wissensquellen bestehen aus der Übertragung von Erfahrungen mit realen Menschen und vielleicht aus Wissen, das über Literatur oder Film angeeignet wurde. Gerade die persönlichen Erfahrungen können Grund für falsche Vorannahmen für den Aufbau einer gemeinsamen Kommunikationsbasis zwischen Mensch und künstlichem Kommunikator sein. Aus diesen Erfahrungen heraus kommt es bei der Interaktion zwischen dem Robotersystem BIRON [Haa04] und einer naiven Person (siehe Kap. 10) daher oft zu Problemen in der Spracherkennung. Einige Probanden gehen dann dazu über, lauter oder langsamer zu sprechen oder auch einzelne Worte zu überartikulieren. Sie übertragen damit die Annahme, dass eine schlecht verstehende Person schwerhörig ist und lauter Sprechen demnach zur Lösung der Kommunikationsschwierigkeiten führt. Für das Robotersystem entstehen jedoch Spracherkennungsfehler ganz anders. Es muss ein Gesicht detektieren, um Sprache interpretieren zu können. Sonst nimmt es an, das seien Störgeräusche aus der Umgebung. Wenn nun der Interaktionspartner lauter spricht, wird das Problem dadurch nicht behoben, sondern kann es sogar noch verschlimmern, weil die Signale womöglich zusätzlich

noch übersteuert sind. Auch plötzliches Langsamsprechen ist für den Spracherkenner deutlich schwieriger. Würde die Person sich stattdessen besser auf den Roboter ausrichten und beim Sprechen nach vorne in die Kamera schauen, so wäre das Verständigungsproblem gelöst. Oftmals ist es nicht möglich, im Vorfeld genaue Vorhersagen über die Interaktion zwischen Mensch und Maschine zu machen. Ein wichtiger Bestandteil bei der Entwicklung realer Systeme sind daher Experimente und Evaluationen, um das System verbessern zu können.

Für eine erfolgreiche Kommunikation ist es essenziell, dass eine kooperative Beziehung zwischen den beteiligten Akteuren aufgebaut wird. Viele Aspekte, die normalerweise während einer Interaktionssituation automatisch angenommen und interpretiert werden, müssen daher vom künstlichen Kommunikationssystem direkt sichtbar gemacht werden. Dabei ist abzuwägen, welche Aspekte implizit bleiben und welche explizit dargestellt werden, um die Natürlichkeit und Intuitivität einer Interaktion nicht zu gefährden und die Kommunikationspartner nicht zu überfordern.

Beispielsweise stellt der Roboter Kismet [Bre99] Emotionen wie Ärger, Ermüdung, Abneigung, Begeisterung, Freude, Interesse und Überraschung dar, indem er verschiedene Gesichtsausdrücke zeigt. Diese Emotionen können unter anderem dazu genutzt werden, einen Kommunikationspartner so zu beeinflussen, dass dieser sich dem Roboter gegenüber möglichst kooperativ verhält, was sich in der Regel positiv auf die Interaktion auswirkt [Bro99]. Kismet reagiert beispielsweise interessiert und fröhlich, falls das Gesicht des Benutzers dem Roboter zugewandt ist und sich der Mensch nur langsam bewegt. Sollte sich der Benutzer zu schnell bewegen, wirkt der Roboter frustriert oder abgeneigt, da es in dem Fall schwieriger für das System ist, den Benutzer zu verfolgen. Zusätzlich verfügt Kismet über eine so genannte *Comfort Zone*: Sollte ein Benutzer zu nah oder zu weit entfernt sein, so kann ihn Kismet bitten, eine für die Sensoren des Roboters günstigere Position einzunehmen [Bre00]. Ein weiterer Roboter, der einen entsprechenden Satz an Emotionen besitzt, ist Pong [Har01]. Ähnlich wie bei Kismet basieren die Reaktionen dieses Systems auf einer Kombination aus Sprecherlokalisation, Spracheingabe und Gesichtsdetektion. Das Robotersystem BIRON [Haa04] besitzt die Möglichkeit, Emotionen mit Hilfe eines Gesichtes auf einem Display darzustellen (siehe Abschnitt 5.5.1) sowie die Interaktion auf Grundlage des Konzeptes der *Joint Attention* [Nag04, Kap04] aufzubauen.

Im Forschungsbereich der Mensch-Maschine-Kommunikation besteht ein großes Interesse am Design und an der Interaktionsgestaltung. Der Designprozess ist fokussiert auf die Gestaltung der äußerlich wahrnehmbaren und erfahrbaren optischen und funktionalen Eigenschaften eines Systems [PB02]. Zusätzlich haben soziale Aspekte einen Einfluss auf den Einsatz computergestützter Systeme. Norman [Nor94] sagt aus, dass im Umgang mit virtuellen Interaktionspartnern gerade die kritischen Punkte Vertrauen, Sicherheit und Erwartungshaltung großen Einfluss auf Erfolg oder Misserfolg „intelligenter“ Systeme haben. In der Praxis zeigen sich meist erst in der Nutzung die Probleme solcher Systeme. Hier ergibt sich die Aufgabe, den Designprozess partizipativ umzusetzen [PB02, Flo97].

Im Gegensatz zur klassischen Mensch-Maschine-Kommunikation, wie z. B. die Fahrzeugführung, Computerapplikationen oder „Computer Supported Cooperative Work“ (CSCW), gibt es bisher nur wenige Studien zur Mensch-Roboter-Kommunikation (siehe z. B. [Wre04a, Hüt03]). Allgemeingültige Aussagen für die Gestaltung der Interaktion können nicht gemacht werden.

Aufgrund der sehr individuellen Gestaltung der Szenarien und der Kontexte, in denen sich die Roboter bewegen (Einrichtungsgegenstände, Handlungskontext, kulturelle Regeln usw.), ist es zudem sehr schwierig, allgemeingültige Kriterien und Regeln für die sinnvolle Gestaltung der Interaktion aufzustellen, da nach der Behavior-Setting-Theorie jede Neuformierung der Umgebung auch jeweils veränderte Verhaltensmuster in den beteiligten Personen weckt. Es ist somit für die Entwickler künstlicher Interaktionspartner schwierig, die Systeme so umzusetzen, dass sie möglichst natürlich handeln. Es ergibt sich zudem die Frage, was *natürlich* für ein künstliches System bedeutet. Die Forschung kann oft nur durch die Erfahrung bei der Entwicklung und vor allem durch die Evaluation ihrer Systeme bei der Interaktion mit naiven Probanden lernen und daraufhin ihre Systeme verbessern. Dazu kann das Wissen über die vorhandenen Gegebenheiten und Strukturen in der Umgebung von Nutzen sein, denn über sie können im Gegensatz zum Verhalten des Benutzers Vorannahmen gemacht werden.

Das Robotersystem BIRON wird auch als ein System verstanden, mit dem verschiedene Interaktions- und Verhaltensmuster der Benutzer beobachtet werden können, und das auf dieser Basis schrittweise verbessert und erweitert wird. Es ist für das Robotersystem von zentraler Bedeutung, dass es eine flexible Architektur besitzt, die erweiterbar und veränderbar ist sowie einen Sprachumfang besitzt, der leicht auf neue Kontexte adaptiert werden kann.

2.3. Situierte Kommunikation

Clancey [Cla97] hat den Ansatz der *Situierten Kognition* vorangetrieben. Der Begriff *Situiertheit* beschreibt die Fähigkeit eines Systems, sich an seine Umgebung anzupassen. Ein situiertes System ist in der Lage, seine Umgebung wahrzunehmen, sie zu manipulieren sowie mit den Kooperationspartnern in der Umgebung zu kommunizieren. Dafür wird die aktuelle Situation möglichst weitgehend als Informationsquelle herangezogen [Lob93b]. Was Menschen wahrnehmen, verstehen und wie sie handeln, hängt in sich zusammen und wird mitbestimmt durch die aktuelle Situation. Dabei setzt sich eine Situation wiederum aus der momentanen Wahrnehmung, dem Szenario, in dem diese Wahrnehmung lokalisiert ist, aus der aktuellen Handlung, aus dem zielgerichteten Plan, aus den Interaktionspartnern und aus der sprachlichen und nicht-sprachlichen Interaktion mit den Partnern zusammen [Lob98].

Die Interaktion zwischen den internen Denkprozessen und der externen Welt ist zentral für den Begriff der Situiertheit. Relevant dabei ist auch der Zugang zu Feedback sowohl über die Interaktion der internen Prozesse untereinander als auch über die Interaktion zwischen den internen Prozessen und der Umgebung. Wissen erhält somit einen dynamischen Aspekt: Es verändert sich kontinuierlich während der Ausführung von Handlungen.

Klassische Mensch-Computer-Dialoge können am ehesten mit Telefongesprächen verglichen werden. Die Dialogpartner können sich nur auf Teile des Gesprächs beziehen, jedoch nicht auf Ereignisse oder Gegenstände in ihrer jeweiligen Umgebung, da der Partner keinen Zugriff auf das Wissen der anderen Person hat. Daraus folgend kann Wissen und die semantische Interpretation ausschließlich aus den Äußerungen beider Dialogpartner generiert werden.

Anders sieht es bei der Kommunikation zwischen Mensch und Roboter-Gefährten aus. Beide Konversationspartner sind in eine gemeinsame Umgebung eingebunden. Sie interagieren in dieser Umgebung, haben Wissen über ihre Umwelt und integrieren es in ihren Dialogen [Hüw04]. Menschen verwenden zwar Sprache als Hauptmodalität, nutzen jedoch auch andere Modalitäten wie Mimik und Gestik. Wissen über Objekte in einer gemeinsamen Umgebung wird ebenfalls vorausgesetzt. Sprache und Gestik bilden häufig eine Einheit, die nur zusammen genommen sinnvoll interpretiert werden kann. Ohne das Szenen-Wissen sind die Äußerungen vielfach ambig oder gar nicht interpretierbar. Äußerungen enthalten häufig Ellipsen, indirekte Sprechakte oder Objekt- und Aktionsreferenzen [Mil97]. Oftmals wird Hintergrundinformation benötigt, um die Äußerung zu verstehen. Deiktische Ausdrücke können nur in der konkreten Situation verstanden werden (z. B. als situierte Objektreferenz). Um in einer realen Welt interagieren zu können, benötigen Robotersysteme verschiedenste Perzeptoren und Fähigkeiten. Sie müssen verstehen können, was der Kommunikationspartner sagt, müssen ihre Umgebung sehen, die Objekte, mit denen sie interagieren, erkennen und Wissen darüber besitzen oder auch neue Objekte kennen lernen. Neben den akustischen Perzepten benötigen sie selbst auch die Fähigkeit, sich mitzuteilen, sie benötigen Wissen über den Sprachgebrauch und Fähigkeiten wie z. B. eine Sprachausgabe.

Die Situation, z. B. die Ansprüche des Benutzers und die Fähigkeiten des Roboters, sowie die Wahrnehmung der Umgebung bestimmen hauptsächlich das Kommunikationsverhalten. Ein komplexes Zusammenspiel zwischen Sprache, Wahrnehmung und Aktion findet statt, daher kann eine korrekte Interpretation des Gesagten nur mit Hilfe von extralinguistischem Wissen generiert werden. Es ist dafür hilfreich, wenn das System Möglichkeiten besitzt, verschiedene Modalitäten miteinander zu verknüpfen, indem es zum Beispiel Hinweise aus der Sprache auf Szeneinformationen aufgreift, und mit visuellen Informationen abgleicht und zusammen abspeichert.

2.4. Spontansprache

Neben der situierten Kommunikation haben mobile Robotersysteme häufig mit den Besonderheiten von Spontansprache zu tun. Sie unterscheidet sich wesentlich von geschriebener Sprache: Oberflächlich betrachtet weist gesprochene Sprache eine eher einfache syntaktische Struktur auf [Mil97]. Oftmals jedoch folgt sie nicht den Regeln der geschriebenen Sprache. Mehr noch, Satz- oder Äußerungsgrenzen sind nicht immer klar auszumachen, da häufig mitten im Satz Pausen auftreten [War94]. Spontansprache beinhaltet beispielsweise unvollständige Äußerungen, Sprecher brechen mitten im Satz ab, verbessern sich und verwenden ungrammatikalische Konstruktionen [Kro00, Lev83, McK98, Pet99]. Satzreparaturen tauchen laut [Hir94] im Korpus des *Air Travel Information System* (ATIS) in etwa 10% der spontansprachlichen Äußerungen auf. Prosodische Informationen spielen eine große Rolle, sie können mitunter den pragmatischen Gehalt einer Äußerung bestimmen (z. B. kann eine Frage nur an der Tonhöhe am Satzende erkennbar sein) [Nöt89]. Im ATIS-Kontext werden prosodische Informationen genutzt, um die Leistung der Spracherkennung zu verbessern [Hir94]. Obwohl das Thema Spontansprache in der

Mensch-Maschine-Kommunikation eine eher untergeordnete Rolle spielt, existieren einige Verarbeitungsmethoden in diesem Bereich (siehe Kap. 4.2), für die eine ausführliche Korpusstudie erstellt wurde. In [McK98] beschreibt McKelvie beispielsweise Disfluenz im englischen Korpus von Wegbeschreibungen. Spontansprachliche Phänomene für Konstruktionsanweisungen im Deutschen werden in [Kro00] ausführlicher beschrieben. Weiteres findet sich in Kapitel 6, das näher auf die Analyse situierter Dialoge und spontansprachlicher Phänomene im Kontext der Mensch-Roboter-Kommunikation eingeht.

2.5. Die Sprechakttheorie und Dialogakttheorie

Ein zentrales Problem bei der Interaktion zwischen Mensch und künstlichem Kommunikationssystem liegt darin, dass die Intentionen und Erwartungen der Kommunikationspartner nicht klar erkennbar sind. Daraus resultieren viele Missverständnisse. Diese Probleme treten sowohl in klassischen Mensch-Maschine-Kommunikationssystemen wie z. B. Nachrichtensystemen als auch in der Mensch-Roboter-Kommunikation auf. Mit Hilfe von Nutzungskonventionen wurde z. B. in Bereichen der computergestützten Kommunikationssysteme (CSCW) versucht, die angemessene Nutzung zu unterstützen [Bro83]. Leider gelang dies bislang nur unzureichend, und so wurde die Sprechakttheorie von Austin [Aus62] durch Winograd und Flores [Win86] populär.

Mit dieser Theorie möchte Austin den Zusammenhang zwischen sprachlicher Äußerung und der Absicht des Sprechers und dessen Wirkung auf eine Handlung erfassen. Sie gibt Formalismen zur Klassifizierung der Zwecke von Äußerungen und erlaubt damit, die Reaktionserwartungen auszudrücken. Dabei wird unterschieden zwischen der Äußerung und der Intention, die der Sprecher damit verbindet. Unterschieden wird zwischen *lokutionären*, *illokutionären* und *perlokutionären* Akten.

- Ein *lokutionärer* Akt ist eine Feststellung, eine Tatsachenbehauptung.
- Ein *illokutionärer* Akt enthält eine beabsichtigte Wirkung auf den Empfänger, man verspricht ihm etwas, man warnt ihn, droht etc.
- Ein *perlokutionärer* Akt enthält die beabsichtigte und erfolgte Wirkung auf den Empfänger, d. h. die Illokution des Senders hat eine Wirkung auf den Empfänger gezeigt.

Eine *lokutionärer* Äußerung kann auf ihren Wahrheitswert überprüft werden. Die vom Sender beabsichtigte Wirkung eines *illokutionärer* Aktes kann sowohl durch sprachliche Äußerungen ausgedrückt werden, indem der Sprecher z. B. sagt, „ich warne dich“ oder „ich verspreche dir“ als auch durch non-verbale Signale, durch die Art, wie er die Äußerung macht. Damit wird eine tatsächliche Handlung ausgeführt, eine Drohung vollstreckt oder ein Versprechen gegeben. Die illokutionäre Bedeutung kann auch durch den verwendeten Sprachmodus (z. B. Imperativ), durch die Betonung, durch Adverbien oder adverbiale Bestimmungen, Konjunktionen, das begleitende Verhalten des Sprechers oder durch die Umstände der Äußerungssituation erschlossen

werden [Aus85]. Auch die Umstände, unter denen etwas gesagt wird, können eine Handlung vollziehen und die Situation verändern, am deutlichsten wird dies am Beispiel der Eheschließung. Beim *perlokutionären* Akt handelt es sich um eine Wirkung, die durch den Sprechakt bewusst hervorgebracht wurde, z. B. ist das Ziel des Überredens, dass der Hörer überredet ist. Jedoch ist umstritten, ob es sich hierbei tatsächlich um einen Akt handelt oder eher um einen Effekt, bei dem fraglich ist, inwiefern diese Effekte überhaupt systematisch erfasst werden können.

Austin beschreibt durch seine Theorie Zusammenhänge zwischen Sprechen und Handeln. Einerseits können Handlungen als Äußerungen verstanden werden, die einen Sprechakt darstellen (z. B. Kopfnicken), andererseits können rein sprachliche Äußerungen, die Wirkung von Handlungen haben, eine Situation verändern. Die Sprechakttheorie verdeutlicht das Zusammenspiel zwischen Handlung und Sprache. Das Verstehen dieser Wechselwirkung ist wichtig für die Umsetzung von Systemen mit kommunikativen und sozialen Fähigkeiten. Erst dann kann eine Verständigung auf höherer Ebene entstehen und Missverständnisse können vermieden werden.

Eine Abwandlung der Sprechakttheorie ist die Dialogakttheorie. Für die Interaktion in einem Dialog zweier Kommunikationspartner, also insbesondere auch in der Kommunikation mit einem künstlichen Kommunikator, trifft dieser Begriff besser auf die Beschreibung von Dialogsequenzen. Im Unterschied zu den illokutionären Akten ist ein Dialogakt nach [Sch95] eine Abstraktion auf pragmatischer Ebene. Er charakterisiert eine Äußerung unabhängig von ihrer grammatikalischen Realisierung und findet in der automatischen Sprachverarbeitung Verwendung. In der Nutzung für Dialogsysteme werden oftmals feinere pragmatische Unterscheidungen von Dialogakten gemacht, als sie in der Sprechakttheorie ausdifferenziert sind. Das Dialogsystem von Brand-Pook [BP99a] beispielsweise nutzt für das Korpus von Konstruktionsanweisungen unter anderem die Dialogakte *Verbindung herstellen, Verbindung lösen, Allgemeine Bauanweisung*. Sie ließen sich auch allgemeiner als ein Dialogakt *Anweisung* beschreiben, jedoch kann das Dialogsystem dann zwischen den genaueren semantischen Informationen nicht mehr unterscheiden. Die genauere Ausdifferenzierung der Dialogakte hängt somit stark vom Kontext und der Nutzung ab.

In natürlichen Kommunikationsszenarien kann jedoch vom Kommunikationspartner nicht erwartet werden, dass er den Dialog- oder Sprechakt explizit übermittelt: Zum einen hat ein naiver Nutzer im Zweifelsfall nicht die Kenntnis über die genaue Bezeichnung seines geäußerten Sprechaktes und zum anderen wäre solch eine Interaktion sehr aufwendig und würde die Natürlichkeit der Interaktion unterbinden. Kommunikationssysteme, die die Sprechakttheorie umsetzen (siehe z. B. [Pri89], wurden daher von den Benutzern eher weniger akzeptiert. An einen künstlichen Interaktionspartner besteht daher der Anspruch, die Absicht des Benutzers direkt aus der Äußerung oder indirekt durch weitere Hinweise aus anderen Modalitäten zu gewinnen.

2.6. Zusammenfassung

Im Kontext der Mensch-Roboter-Kommunikation ist es nicht sinnvoll, Sprache ganz unbeeinflusst von anderen Prozessen zu betrachten. Es findet immer ein Wechselspiel vieler Einflüsse statt, die den gesamten Verlauf eines Dialoges beeinflussen und die die Qualität der Kommunikation und der im Kontext verwendeten Äußerungen mitbestimmen. Daher ist ein Roboter mit sozialen und kommunikativen Fähigkeiten ein komplexes System, auf das viele verschiedene Aspekte Einfluss nehmen, die wiederum den Interaktionsprozess beeinflussen.

Die wichtigsten Konsequenzen, die man aus den theoretischen Überlegungen für die Entwicklung von sozialen Robotersystemen ziehen kann, sind hier noch einmal zusammengefasst:

- Eine Äußerung kann auch immer als Handlung gesehen werden. Eine Handlung ist auch eine kommunikative Mitteilung.
- Die Beziehung der Kommunikationspartner beeinflusst das Kommunikationsverhalten, es trägt zur gelungenen oder missglückten Kommunikation bei.
- Vorannahmen aufgrund von Erfahrungen bei der natürlichen Kommunikation mit anderen Personen kann die Mensch-Roboter-Kommunikation sowohl unterstützen als auch wesentlich stören.
- Die Umgebung fließt immer in die Kommunikation mit ein. In realen Szenarien beeinflusst das die Struktur der Äußerungen wesentlich.
- Die direkte Interaktion zwischen Mensch und Roboter in einer gemeinsamen Umgebung wird durch die Besonderheiten der Spontansprache geprägt.
- Die Szenarien, in denen Roboter agieren, sind oftmals so verschieden, dass jeder Aufbau für sich konzipiert werden muss. Allgemeine Designkriterien können nicht aufgestellt werden, sie variieren je nach Kontext.
- Wissen über den Dialogakt einer Äußerung erleichtert das Erkennen der Absicht des Benutzers und kann somit die Kommunikationsfähigkeiten des Roboters unterstützen.

3. Ansätze zum Sprachverstehen

In der Entwicklung von computergestützten Alltagsgeräten (z. B. PDAs, Handys und Automobilen) spielt vor allem der Informationsaustausch mittels Sprache eine immer größere Rolle. Der Trend wird sich vermutlich in den nächsten Jahren fortsetzen. Die automatische Verarbeitung von Sprache nimmt daher einen großen Stellenwert in der Entwicklung so genannter „intelligenter Systeme“ ein. In komplexeren Interaktionssystemen ist nicht nur die Darstellung der reinen Äußerung, sondern sind vor allem die semantische Bedeutung und Informationen zur pragmatischen Verwendung relevant. Verstehen bedeutet in einem künstlichen System, für die sprachlichen Äußerungen eine Repräsentation der Bedeutung zu erstellen. Es geht dabei weniger um das kognitive Verstehen oder Begreifen, sondern um das Bereitstellen semantischer und ggf. pragmatischer Informationen, die das Gesamtsystem für die Erledigung seiner Aufgaben benötigt.

Im Folgenden werden zunächst zentrale Begriffe und Theorien der Linguistik vorgestellt, die für diese Arbeit von großer Bedeutung sind. Ebenfalls werden verschiedene Repräsentationsformalismen für die Speicherung sprachlicher Informationen erläutert. Da es in der Linguistik viele unterschiedliche Definitionen und Theorien zur Verarbeitung menschlicher Sprache gibt, die für den Kontext der Mensch-Roboter-Kommunikation jedoch nicht zur Klärung beitragen, werden hier nur allgemein die zentralen Aspekte beschrieben. Im letzten Abschnitt dieses Kapitels werden Ansätze zur Verarbeitung von Sprache, die speziell in Dialogsystemen für die Mensch-Maschine-Kommunikation zum Einsatz kommen, exemplarisch vorgestellt. Sie berücksichtigen jeweils verschiedene Aspekte, die für eine komplexere Interaktion mit einem sozialen künstlichen System zum Tragen kommen. Meist greifen Repräsentationsformalismen und Verarbeitungsmechanismen ineinander. Wie die gesamte Verarbeitung von Sprache in einem System funktioniert, wird in Kapitel 4 exemplarisch an verschiedenen Dialogsystemen ausgeführt.

3.1. Theoretische Grundlagen

In diesem Abschnitt werden grundlegende Begriffe beschrieben, die zum Verständnis der nachfolgenden Ausführungen hilfreich sind. Dabei sind die Definition hier möglichst allgemein gehalten, wobei zusätzlich der Bezug zur Sprachverarbeitung erläutert wird.

3.1.1. Begriffsdefinitionen

Die *Syntax* beschreibt die Relationen sprachlicher Zeichen zueinander [Bra99]. Dabei übernehmen unterschiedliche Wortarten bestimmte Funktionen. Mit der Syntaxanalyse lassen sich spezielle Einheiten innerhalb eines Satzes ausmachen, komplexe Strukturen mit Hilfe von Regeln beschreiben.

In der *Semantik* geht es um die Bedeutung der Äußerungen. Jeder lexikalischen Einheit wird eine Bedeutung zugeordnet und auch die Gesamtbedeutung eines Satzes wird ermittelt [Bra99]. Sie beschreibt ebenfalls, wie sprachliche Ausdrücke auf die Außenwelt bezogen sind. Um einzelnen Wörtern und ganzen Äußerungen eine Bedeutung zuordnen zu können, muss eine Menge Wissen vorhanden sein: Weltwissen, Diskurswissen und Kontextwissen. Das semantische Wissen stellt einen Bezug zwischen intensionaler Bedeutung und der realen Welt dar, mit dessen Hilfe Befehle oder Anweisungen für ein kommunikatives System generiert werden können. In sprachverarbeitenden Systemen vermischt sich das Wissen über Syntax und Semantik oftmals. Analyseprozesse verbinden häufig in einem einzigen Schritt beide Elemente der Sprache. Dies ist sinnvoll, da letztendlich die semantischen Informationen für den Interaktionsprozess eines künstlichen Kommunikators ausschlaggebend sind.

Die *Pragmatik* beschreibt Sinn und Funktion einer Äußerung. Sie erklärt, wie ein bedeutungsvoller sprachlicher Ausdruck in einer Situation zur Realisierung einer bestimmten Sprecherintention gebraucht wird ([Bra99] S. 313). Informationen aus der Sprache reichen nicht vollständig aus, um pragmatisches Wissen generieren zu können. Kontextwissen, wie z. B. die Sprecherintention oder Wissen über die Umgebung, fließt ein. Pragmatisches Wissen kann ein Dialogsystem daher nur in Verbindung mit den zur Verfügung stehenden Kontextinformationen generieren. Der Dialogakt der Äußerung ist eine wichtige Stütze für den Erwerb interner Pläne und der Dialogfähigkeiten eines kommunikativen Systems. Er kann jedoch meist schon unter Zuhilfenahme zusätzlicher Informationen über die Aufgabenstellung des Systems aus der Äußerung des Interaktionspartners gewonnen werden.

3.1.2. Theta-Rollen-Theorie

Die Theta-Rollen-Theorie ist eine der grundlegenden Theorien der Semantik, die eine Grundannahme für viele Ansätze zur Sprachverarbeitung und Modelle der Linguistik ist sowie auch für die Konzeption der *Situierten Semantischen Einheiten* in Kapitel 8.

Die Grundidee ist, dass in jedem Satz einzelne Argumente bestimmte semantische Rollen übernehmen. Dabei versteht man unter semantischen Rollen eine *Bedeutungsfunktion* eines Satzteils in Bezug auf den ganzen Satz. Man sagt auch, die Argumente realisieren eine bestimmte thematische Rolle oder *Theta-Rolle*. Dabei existieren eine Reihe von verschiedenen Rollen in einem Satz. So hat z. B. ein Argument, meist als Subjekt (im Aktiv), die Aufgabe, den Täter oder AGENS einer Handlung darzustellen, ein anderes Argument stellt den PATIENS (der die Handlung erfährt) dar. Welche Rollen vergeben werden und wer welche Rolle übernimmt, hängt vom

Kontext oder „Drehbuch“ für den Satz ab. Meist bestimmt das Verb die thematischen Rollen des Satzes. Deshalb sagt man, das Verb theta-markiert[1] seine Argumente. Weitere Rollen sind beispielsweise INSTRUMENT, GOAL, LOCATION, SOURCE. Dabei ist die Anzahl der semantischen Rollen nicht festgelegt und variiert je nach Variante und Gebrauch.

Die Idee der Vergabe von semantischen Rollen geht auf Fillmore [Fil68] zurück, der AGENS, PATIENS, etc. als Tiefenkasus bezeichnet hatte. Neben Fillmores Arbeiten spielen auch die Ansätze von Gruber [Gru67] und Jackendoff [Jac72] eine wichtige Rolle, sie sprechen dabei jedoch von „Kasusrollen“.

Die Theta-Rollen-Theorie kommt in vielen Grammatikformalismen zur Anwendung. Die Rollen werden dazu verwendet, semantische Grundverhältnisse innerhalb von Sätzen zu beschreiben. Gemeinsam ist diesen Bemühungen der Versuch, syntaktische Größen zur Beschreibung von semantischen Verhältnissen einzusetzen. Ein Bestreben ist u. a. Syntax und Semantik in einem einzigen Grammatikmodell zu erfassen. In einigen Grammatiktheorien bilden sie die Schnittstelle zwischen Semantik und Morpho-Syntax. Die Theorie der semantischen Rollen ist beispielsweise wesentlicher Bestandteil der *Kasusgrammatik* [Fil68], der *Generativen Grammatik* von Chomsky [Cho81] sowie der *Funktional Grammar* [Dik91]. Die semantischen Rollen können in vielen Fällen vorhandene Strukturen aufklären und daher für die Analyse von Äußerungen genutzt werden.

Die Grundidee der Theta-Rollen-Theorie fließt ebenfalls in den semantischen Konzepten der *Situierten Semantischen Einheiten* (siehe Kap. 8) ein, die für die Repräsentation des Wissens für die Interaktion zwischen Mensch und Roboter konzipiert wurden. Jedoch ist eine strikte Zuordnung von syntaktischer Funktion zur Rolle in Äußerungen, die einen Handlungsbezug beinhalten können, nicht unbedingt sinnvoll. Eine Person kann eine Anweisung an den Roboter auf unterschiedliche Weise geben, z. B. durch „Drehung rechts“ oder „drehe dich nach rechts“ und meint damit jeweils dieselbe Aufforderung, sich nach rechts zu drehen (siehe auch Kap. 10). Syntaktisch ist dabei nur im zweiten Beispiel ein Verb beteiligt, das Wort „Drehung“ nimmt im ersten Fall jedoch dieselbe Rolle wie das Verb an.

3.2. Repräsentationsformalismen

Es gibt viele Möglichkeiten, semantisches Wissen darzustellen. Je nach Aufgabenstellung werden unterschiedliche Schwerpunkte gesetzt: welche Informationen bereitgestellt werden, in welchem Format die Daten gespeichert sind und wie der Zugriff darauf umgesetzt ist. In diesem Abschnitt werden prominente Repräsentationsformalismen zur Beschreibung semantischer Informationen vorgestellt. Sie bilden alternative Ansätze zu den *SSUs*, die in Kapitel 8 erläutert werden. Verschiedene Sprachverarbeitungssysteme nutzen unterschiedliche Repräsentationsformalismen für ihre jeweilige Aufgabenstellung. Dabei spielt neben den theoretischen Überlegungen der linguistischen Modelle, in deren Rahmen sie zum Einsatz kommen, auch die Problem-

[1] theta vom griechischen Buchstaben θ für thematisch

stellung der Aufgabe eine wichtige Rolle, z. B. der Umfang des Wortschatzes oder der Kontext des Interpretationsbereiches. Nach speziellen Kriterien ausgewählte Dialogsysteme, in denen semantische Informationen genutzt werden, werden im nachfolgenden Kapitel 4 beschrieben.

3.2.1. Formale Semantik

Die formale Logik (oder auch formale Semantik) beschreibt die semantischen Relationen zwischen Sätzen in einer Sprache. Dieser Ansatz basiert auf dem Begriff der *Wahrheit*. Danach wird die Bedeutung eines Satzes aufgrund seines Wahrheitsgehalt bestimmt. Da in vielen Lehrbüchern Logikansätze ausführlich beschrieben sind (siehe z. B. [Sae97, Rus95]), wird hier auf die genaue Beschreibung des Formalismus verzichtet und nur zentrale Aspekte für die Verarbeitung gesprochener Sprache aufgegriffen.

Wissensbasierte Systeme wie z.B. Expertensysteme oder „Question-Answering“ Systeme nutzen vorwiegend formal-logische Repräsentationen der Prädikatenlogik. Sie lassen sich leicht in Inferenzmechanismen einbinden und können vom System verarbeitet werden. Es werden mit ihrer Hilfe die Beziehungen zwischen Sätzen hergestellt und der Wahrheitsgehalt einer Äußerung aus dem vorhandenen Wissen bestimmt. Somit kann das System auf Fragen vom Anwender entsprechend antworten.

Für andere Aufgabenbereiche sind Logikformalismen wie die Prädikatenlogik weniger geeignet. Sie sind zum einen nicht reich und flexibel genug, um die Bedeutungsstrukturen natürlicher Sprache abzubilden. Logikformalismen bilden einen Wahrheitsgehalt ab, viele Äußerungen besitzen jedoch keinen Wahrheitsgehalt. Es ist unklar, welche logischen Beschreibungen z. B. eine Anweisung wie „komm her“ oder eine soziale Interaktion wie „Einen schönen guten Tag!“ abbilden. In vielen Kontexten der Mensch-Maschine-Kommunikation ist der Handlungsakt oder der soziale Aspekt ausschlaggebend und weniger der Wahrheitsgehalt. Sprachliche Ausdrücke können zudem mehrdeutig sein, was durch die Logik nicht abgebildet werden kann. Kontextinformationen müssten bei der Auflösung von Mehrdeutigkeiten berücksichtigt und entsprechend modelliert werden, was die Prädikatenlogik nicht leistet. [2]

3.2.2. Semantische Netze

Mit semantischen Netzen wird linguistisches Wissen innerhalb eines Netzwerkes repräsentiert. Sie gehen auf Quillian [Qui68] zurück, der sie als ein einfaches Modell des menschlichen Gedächtnisses eingeführt hat. Die Grundelemente sind Knoten und Kanten, die in einem gerichteten Graphen eingebettet sind. Dabei werden Informationen über allgemeine Begriffe (Objekte, Ereignisse etc.) in Knoten und Beziehungen zwischen diesen Begriffen in Kanten repräsentiert. So kann z. B. elementares Wissen über einfache Sachverhalte definiert werden. Abbildung 3.1 stellt ein einfaches Beispiel aus der Sprachverarbeitung dar.

[2] Komplexere Logikformalismen (z. B. die Modallogik) lösen verschiedene Probleme, sind jedoch nicht mehr semi-entscheidbar und daher kaum für die Sprachverarbeitung in Echtzeit-Systemen geeignet.

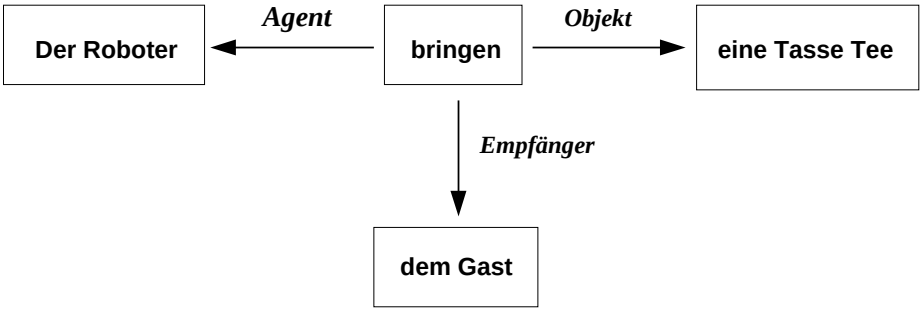


Abbildung 3.1.: Ein einfaches semantisches Netz für die Äußerung: „Der Roboter bringt dem Gast eine Tasse Tee“.

Neben der Darstellung von Beziehungen innerhalb eines Satzes können auch Verflechtungen eines Satzes mit anderen Sätzen beschrieben werden oder ganze Handlungsabläufe zueinander in Bezug gebracht werden. Es können überdies auch Beziehungen zwischen Begriffen und real existierenden Instanzen abgebildet werden, wie in Abbildung 3.2 dargestellt. In dem Netz werden ganz spezielle Kanten wie *mag* oder *kocht* verwendet. Semantische Netze bieten also die Möglichkeit, Beziehungen zwischen Begriffen oder Entitäten durch die Markierung einer Kante auszudrücken. Die Kante wird demnach mit einer *Rolle* belegt. Dabei fußt die Modellierung der semantischen Beziehungen auf der Beschreibung der *semantischen Rollen* nach Fillmore [Fil68] (siehe Abschnitt 3.1.2).

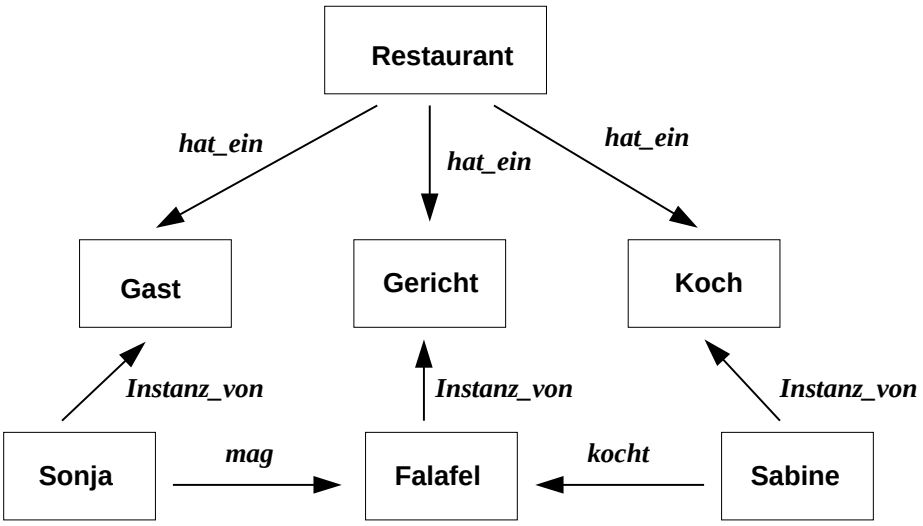


Abbildung 3.2.: Ein semantisches Netz mit Vererbungs- und Rollenbeziehungen.

Obwohl semantische Netze viele positive Eigenschaften für die Wissensrepräsentation bieten, stellen sich einige Probleme ein, die sich vor allem aus der Freiheit der Definition der Knoten und Kanten ergeben. Da die Semantik der Zuordnung frei definierbar ist, lassen sich auch keine allgemeinen Inferenz- oder Ableitungsregeln zur Wissensnutzung aufstellen und verschiedene

Problemstellungen können somit nicht gelöst werden. Aus diesem Grund wurden Formalismen wie z. B. die CD-Theorie [Sha72] oder ERNEST [Mas94] entwickelt, die auf dem Grundkonzept fußen, jedoch zusätzliche Möglichkeiten zur Problemlösung bieten.

Eine Abwandlung der semantischen Netze bildet die Conceptual Dependency Theorie (CD-Theorie) nach Schank [Sha72]. Die Theorie ist in viele Sprachverarbeitungssysteme (z. B. in SAM [Sch90]) eingegangen. Sie legt ebenfalls ein semantisches Netz zugrunde und unterstützt die Theorie der Inferenzbildung. Knoten und Kanten werden durch Konzeptklassen festgelegt. Dabei bilden Aktionen, Objekte, Aktionsmodifikatoren und Objektmodifikatoren die Grundkonzepte der CD-Theorie, wobei alle Handlungsverben durch Aktionen ausgedrückt werden. Mit Hilfe von Abhängigkeitsrelationen zwischen den Grundkonzepten können verschiedene Situationen beschrieben werden. Ebenfalls fließen Informationen der Kasustheorie nach Fillmore [Fil68] ein. Dafür nutzt man das Wissen, dass Verben eine bestimmte Anzahl an Argumenten „besitzen“. Beziehungen zwischen den Satzbestandteilen werden als Netzbeziehungen ausgedrückt (siehe auch Abb. 3.1). Vorteile bei diesem Ansatz sind zum einen die einfache Inferenzbildung, eine eindeutige semantische Darstellung, Verminderung von Mehrdeutigkeiten sowie die weitestgehende Sprachunabhängigkeit. Problematisch ist jedoch, dass die Reduzierung aller Handlungen auf 12 Aktionen zu starkem Informationsverlust führt. Einige Äußerungen können damit gar nicht oder nur sehr umständlich ausgedrückt werden. Beispielsweise wird der Satz “Das Messer fällt auf den Boden.” beschrieben als “Die Schwerkraft befördert das Messer auf den Boden.”

Eine der ersten Implementierungen semantischer Netze im Rahmen einer natürlichsprachlichen Anwendung war das geographische Lehrsystem SCHOLAR [Car70], das Fragen über den südamerikanischen Kontinent beantworten und auch selber stellen konnte. Das Sprachverarbeitungssystem EVAR [Mas94] zur Zugauskunft verwendet ebenfalls semantische Netze zur Wissensrepräsentation und die speziell entwickelte Repräsentationssprache ERNEST. Das System ist für die Verarbeitung kontinuierlich gesprochener Sprache in Telefonqualität ausgelegt. Dabei verwendet ERNEST verschiedene Kantentypen, um sowohl Vererbung, Instantiierung und Teilbeziehungen repräsentieren zu können. Berücksichtigt wird ebenfalls, dass Informationen, die aus einer realen Umgebung gewonnen werden, mit einer gewissen Unsicherheit behaftet sind. Daher wurde eine Bewertungsstruktur eingebunden, die dem Rechnung trägt. Im Rahmen einer Sprachverstehenskomponente in einem Konstruktionsszenario [BP99a] wurde ERNEST für den Kontext von Handlungsanweisungen an einen Konstruktionsroboter erweitert.

Für klar definierte Aufgabenbereiche ist der Einsatz von semantischen Netzen sinnvoll. Problematisch ist allerdings die Entwicklung großer Wissensbasen und vor allem die flexible Handhabung von Informationen. Eine nachträgliche Änderung oder Erweiterung der Netzwerke ist aufwendig und kann zu Inkonsistenzen führen. Daher sind semantische Netze für den Einsatz in kommunikativen Robotersystemen, deren Sprachdaten flexibel angepasst werden müssen, nur sehr bedingt geeignet.

gehen	fliegen
⟨Fortbewegung⟩	⟨Fortbewegung⟩
⟨auf dem Boden⟩	⟨in der Luft⟩
⟨mit Beinen⟩	⟨mit Flügeln⟩

Tabelle 3.1.: Merkmale der Bewegungsverben gehen und fliegen

rennen	schleichen
⟨Fortbewegung⟩	⟨Fortbewegung⟩
⟨auf dem Boden⟩	⟨auf dem Boden⟩
⟨mit Beinen⟩	⟨mit Beinen⟩
⟨schnell⟩	⟨langsam⟩

Tabelle 3.2.: Merkmale der Bewegungsverben rennen und schleichen

3.2.3. Merkmalsbasierte Semantik oder Komponentialsemantik

In der Semantik besteht allgemein das Problem, Sprache mittels sprachlicher Mittel beschreiben zu müssen. Daher wird in vielen Bereichen eine Art Meta-Sprache eingeführt. Die Idee bei der Merkmalsbasierten Semantik ist, dass sich die Bedeutung von Wörtern und Sätzen durch eine Reihe von distinktiven Merkmalen und mittels primitiver Funktionen auf diese Merkmale beschreiben lassen[3]. Die Bedeutungen von Morphemen oder Wörtern sind nach der zugrunde liegenden Annahme zusammengesetzte Konstrukte von semantischen Primitiven oder Komponenten. Beispielsweise lassen sich Gegenstände oder Lebewesen dadurch kennzeichnen: Ein Fisch ist belebt, besitzt Flossen, lebt im Wasser und hat Kiemen. Ebenso sind nach der Idee Verben semantisch aus verschiedenen Komponenten zusammengesetzt wie in Tabelle 3.1. Mit Hilfe der distinktiven Merkmale lassen sich zudem minimale Bedeutungsunterschiede darstellen (siehe Tabelle 3.2). Die semantischen Konstrukte können mittels Attribut-Wert-Strukturen oder mit Matrizen beschrieben werden. In vielen Grammatiktheorien finden sich die merkmalsbasierten Beschreibungen im Lexikon, z. B. in der *Generalisierten Phrasenstrukturgrammatik* (GPSG) [Gaz85], der *Head-driven Phrase Structure Grammar* HPSG [Pol94] oder der *Lexikalisch-funktionalen Grammatik* (LFG) [Bre01].

In Katz und Fodor [Kat63] ist die Verbindung zwischen Syntax und Semantik wesentlich. Wie auch syntaktische Regeln sollen semantische Regeln rekursiv sein. Ebenfalls ist die Bedeutung von Sätzen kompositional. Die Art, wie die Wörter miteinander verbunden sind, bestimmt die Bedeutung von Sätzen. Dafür gibt es neben lexikalischen Einheiten mit einer semantischen Repräsentation auch Projektionsregeln, die aus den lexikalischen Einträgen die Satzbedeutung generieren. Die Theorie der *Conceptual Dependency* nach Schank [Sha72] ist eine Variante der Merkmalssemantik (siehe Absatz 3.2.2). Schank bestimmt Satzbedeutungen mit Hilfe primitiver Akte und hierarchischer semantischer Merkmale.

[3] Eine Einführung findet sich in [Sae97] und [Sch97]

Jackendoffs *Konzeptuelle Semantik* [Jac90] ist ebenfalls eine Form der Merkmalssemantik. Nach dieser Theorie sind alle komplexen Wortbedeutungen aus primitiven Konzepten und Kombinationsregeln zusammengesetzt. Demnach sind semantische Primitive oder Konzepte abstrakte Kategorien. Jedes Konzept gehört zu einem bestimmten Typ, wie z. B. zum Typ der Gegenstände und Lebewesen THING, zum Typ der Ereignisse und Aktionen EVENT, zum Typ der situativen Zustände STATE oder zum Typ der Lokationen POSITION. Weitere Typen sind ACTION, PATH, PROPERTY, AMOUNT. Zusätzlich zu den Typen gibt es primitive Abbildungsfunktionen wie z. B. GO, STAY, FROM, VIA, CAUSE, AT mit denen die konzeptuellen Strukturen aufeinander abgebildet werden. In dem Beispielsatz weist AT der Struktur THING eine bestimmte POSITION zu.

(1) $[_{Event}\ GO\ ([_{Thing}\ JOHN][_{Path}\ TO\ ([_{Position}\ AT\ ([_{Thing}\ SCHOOL])])])]$[4]

Lobin erweitert in [Lob98] die Konzeptuelle Semantik, um Handlungsanweisungen zu modellieren und einen Ansatz für eine angemessene Verarbeitung bereitzustellen. Dabei wird von einem zweistufigen Verarbeitungsansatz ausgegangen, der eine linguistisch motivierte semantische Repräsentation und eine planungsnahe Repräsentation beinhaltet. Die planungsnahe Ebene ist die der Aktionsschemata, sie bilden den Rahmen einer Aktionsbeschreibung. Die zentrale Aufgabe der Aktionsschemata ist die Dekomposition der komplexen Aktionen in Sequenzen von Basisaktionen. Die resultierenden Informationseinheiten eines Aktionsschemas können in so genannte interne Sensorinformationen überführt werden, die die Basis der Kommunikation zwischen deliberativem System und Behavioursystem bilden.

3.2.4. Frame-Semantik und FrameNet

Das Frame-Konzept nach [Fil76] stellt semantische Beziehungen zwischen verschiedenen Satzelementen dar. Ein Frame beschreibt eine abstrakte Situation mit den teilnehmenden Partizipanten und Requisiten. Diese Teilnehmer und Requisiten heißen *Frame-Elemente*, kurz FEs. Dabei gibt es in einem Satz besondere Wörter, die einen Frame einführen oder evozieren können. Das Konzept basiert auf der Theorie von semantischen Rollen (siehe Kap. 3.1.2), genauer der Kasustheorie nach Fillmore [Fil68]. Ähnlich wie bei der Theta-Rollen-Theorie gibt es verschiedene Rollen, die den Wörtern einer Äußerung in einem bestimmten Kontext zugeschrieben werden. Die Rollen sind hier die Frame-Elemente. Sie beschreiben die semantischen Argumente von Verben (und einigen Nomina und Adjektiven) und ihre syntaktische Umsetzung.

Das Projekt FrameNet [Bak98, Fil01] baut auf dem Prinzip der Frames auf. Ziel ist, ein Lexikon für das Englische aufzubauen, in dem für jeden Eintrag eine semantische Frame-Beschreibung gegeben wird. Die FrameNet-Theorie basiert sowohl auf dem theoretischen Frame-Konzept als auch auf empirischen Datenerhebungen. Dafür wurden Sätze des *British National Corpus* und des *LDC North American Newswire Corpus* verwendet und die Wörter mit dem FrameNet-Annotationstool analysiert. Bei den FrameNet-Daten handelt es sich um geschriebene Nachrichten-Texte, die ausschließlich aus grammatikalisch korrekten Sätzen bestehen.

[4] aus [Lob98] S. 82

In dem Projekt werden sowohl lexikalische Informationen als auch semantische Daten abgebildet. Eine *Lexical Unit* ist eine Paar-Struktur eines Wortes und einer in Beziehung stehenden Bedeutung, einem semantischen Frame, genannt *Frame-Element* (FE). In jedem Satz wird mittels syntaktischer Merkmale ein Ziel oder *target* definiert, das einem Wort im Satz entspricht. Entsprechende Label für die Wörter oder Phrasen werden für den speziellen Satz bereitgestellt, der auch als eine Instanz des Frames angesehen werden kann.

In dem Lexikon sind englische Verben, Nomina und Adjektive enthalten, die Frames einführen können. Die Struktur eines lexikalischen Eintrags *Lexical Unit* besteht aus mehreren Teilen: dem Namen oder Kopf des Lexems, dem zugehörigen Frame, einer Definition, einer Liste von FE-Realisationen, einer Liste von Valenz-Mustern (mögliche syntaktische Realisierung) sowie annotierten Beispielsätzen aus dem *British National Corpus*.

Die Frame-Elemente sind Realisierungen von Strukturen, die Annotationen von allen in dem Korpus enthaltenen Sätzen beschreiben. Zu jedem Frame existiert eine Beschreibung der Situation, die der Frame repräsentiert. Ein FE wird angeführt von einem Namen (z. B. Patient) und enthält Informationen über die grammatikalische Funktion (z. B. Objekt oder Modifier) sowie Informationen über den Phrasentyp (z. B. Nominalphrase) der jeweilig zugehörigen Sätze. Die folgenden Beispiele sind Auszüge aus dem FrameNet-Datensatz. Sie illustrieren die Relationen des Frames *JUDGEMENT* innerhalb eines Satzes, das vom Verb „blaming“ evoziert wird (aus [Gil00]).

(2) $[_{Judge} She]$ **blames** $[_{Evaluee}\ the\ Government]$ $[_{Reason}\ for\ failing\ to\ do\ enough\ to\ help]$.

(3) $Holman\ would\ characterize\ this\ as$ **blaming** $[_{Evaluee}\ the\ poor]$.

Die Daten von FrameNet dienen vorrangig der Konzeption eines Lexikons mit semantischen Frame-Beschreibungen, sie sind jedoch nicht mit dem Ziel der Bereitstellung eines Lexikons für die automatische Analyse von Sprache im Kontext der Mensch-Maschine-Interaktion entwickelt worden. Die Frames beschreiben Situationen und repräsentieren nicht alle Informationen einer linguistischen Einheit. Beispielsweise werden Pronomen, Fragewörter usw. nicht abgebildet. Ebenso sind Negationen, vorhergegangene Informationen wie z. B. „decke diesen Tisch genauso wie den vorherigen“ oder Hilfsverben und Tempora nicht durch Frames abgebildet. Sie müsste man durch ein eigenes Modul verarbeiten. Informationsextraktion in bestimmten Kontexten ist jedoch möglich. Jurafsky nutzt die Daten, um Nachrichtenartikel zu analysieren und Wörtern automatisch zugehörige Rollen zuzuordnen [Gil00].

Ebenso wie FrameNet ein Lexikon für das Englische erstellt, bemühen sich einige andere Forschergruppen um die Erstellung in anderen Sprachen. Das Projekt Salsa [Erk03] beispielsweise beschäftigt sich mit der Erstellung lexikalischer Ressourcen für das Deutsche.

WordNet ist ebenfalls ein Projekt, das zum Ziel hat, eine lexikalische Datenbank zu erstellen. Dabei werden semantische Beziehungen zwischen Wörtern durch ein Netzwerk semantischer Informationen abgebildet [Fel99].

3.3. Verarbeitungsmechanismen

Für die Verarbeitung von Sprache existieren eine Vielzahl von Konzepten und Methoden, von denen, aufgrund der Komplexität des Themas, hier nur einige kurz skizziert werden. Die Konzepte und Modelle werden nicht nach theoretischen Überlegungen bewertet, sondern rein nach deren Einsatzmöglichkeiten in der Mensch-Roboter-Kommunikation. Zusätzlich werden in Kapitel 4 einige Verfahren ausführlicher beschrieben, die in verschiedenen Dialogsystemen zum Einsatz kommen.

Je nach Aufgabenstellung können Systeme sehr unterschiedlichen Herausforderungen gegenüberstehen. Daher sind die einzelnen Verarbeitungskonzepte auch unterschiedlich gut für spezielle Problemstellungen geeignet. Nicht alle Verfahren können z. B. in kommunikative Robotersysteme eingesetzt werden. Dabei ergeben sich für den Einsatz spezieller Sprachverarbeitungsmethoden u. a. folgende Fragestellungen: Muss ein System zeitnah reagieren können? Wie komplex ist das Korpus? Welche Art von Eingaben sind zu erwarten? Liegen die Daten in schriftlicher Form vor oder können die Daten möglicherweise fehlerbehaftet sein (z. B. bei der Nutzung einer Spracherkennung)? In einem zeitkritischen System sollte kein Verfahren eingesetzt werden, das eine exponentielle Laufzeit besitzt. Ebenso benötigt ein System, das einfache Anweisungen erhält, keine komplexen Analysemethoden. Die besonderen Anforderungen für die Verarbeitung von Äußerungen im Kontext mobiler Roboter sind in Kapitel 7 dargelegt.

Eine einfache, aber in der Mensch-Maschine-Interaktion gängige Methode ist die Analyse anhand von Schlüsselworten oder nach dem *Pattern-Matching* Verfahren. Dort werden in der Eingabe bestimmte Phrasen oder Schlüsselwörter gesucht, die direkt in ein semantisches Schema überführt werden können. Das System ELIZA [Wei66] ist eines der ersten Systeme, das mit dieser recht einfachen Methode beachtliche Interaktionsfähigkeiten besitzt.

Viele Sprachverarbeitungsysteme setzen Parser ein, die mit Hilfe unterschiedlicher Grammatikformalismen linguistische Einheiten verarbeiten können. Dabei wird überprüft, ob ein Satz syntaktisch korrekt ist, zusätzlich werden den Satzbestandteilen syntaktische Funktionen zugeordnet. Eine der ältesten Theorien ist die der *Transformationalen* oder *Generativen Grammatik* [Cho72]. Neuere Ansätze sind die *Lexikalisch-funktionale Grammatik* (LFG) [Bre01] oder die *Head-driven Phrase Structure Grammar* (HPSG) [Pol94]. Sie arbeiten auf Basis des Unifikationsmechanismus und nutzen zur Modellierung der einzelnen Wörter Merkmalsstrukturen, die die syntaktischen und semantischen Eigenschaften modellieren. Das heißt, neben den grammatikalischen Regeln enthalten die lexikalischen Einträge wichtige Informationen, die während der Unifikation verarbeitet werden. Ein Problem gerade bei den generativen Grammatiken besteht darin, dass Sprachen mit freier Konstituentenreihenfolge kaum berücksichtigt wurden und nur sehr aufwendig zu realisieren sind (z. B. HPSG für das Deutsche in [Mül99]). Generell besteht bei Satzgrammatiken wie der LFG oder HPSG die Gefahr, dass das Parsen einer Äußerung wegen eines möglicherweise kleinen Erkennungsfehlers komplett fehlschlägt. Auch folgt gerade gesprochene Sprache nicht immer den grammatikalischen Regeln der geschriebenen Sprache.

Für Sprachen mit freier Wortstellung bietet sich eher die *Dependenzgrammatik* an. Die Dependenzgrammatik (siehe auch [Mel88, Lob93a]) beschreibt Abhängigkeiten zwischen Wörtern eines Satzes. Dabei regiert ein Wort (Regens) ein oder mehrere Wörter (Dependentien oder Aktanten). Dependentien können wiederum Regens sein, dann bilden sie sogenannte Satelliten des vorangehenden Regens. Dabei wird bei der strukturellen Analyse vom Verb als Wurzel der Struktur ausgegangen, es erhält die Sonderstellung des Regens. Die Valenz (oder Wertigkeit) eines Verbs bestimmt die Anzahl möglicher Ergänzungen. Für jedes Verb können neben der Anzahl der Ergänzungen auch die Art der Ergänzungen (z. B. Substantiv im Akkusativ, Dativ etc.) und die semantischen Merkmale (z. B. Menschen oder Dinge als Dependentien) festgelegt werden. Im Gegensatz zur Generativen Grammatik geht die Dependenzgrammatik nicht von festen Satzmustern aus. Dennoch werden den Wörtern Kategorien ähnlich den grammatischen Relationen zugeordnet, die mögliche Verb-Ergänzungen darstellen.

Eine Sonderform der Generativen Grammatik ist die Kasusgrammatik [Fil68], die in einigen Aspekten der Dependenzgrammatik ähnelt. Nach diesem Modell besitzt ein Satz eine Oberflächen- und eine Tiefenstruktur, wobei die Tiefenstruktur durch semantische Kasusrollen eines Satzes charakterisiert wird und dieser Tiefenkasus die Oberflächenstruktur als syntaktische Funktionen realisiert. Bei dem Tiefenkasus handelt es sich um universelle semantische Rollen, die den „Mitspielern“ oder Aktanten im Satz zugeteilt werden können. Sie gehen auf die Kasusrollentheorie zurück (siehe Absatz 3.1.2). Dabei selegiert jedes Verb (und Adjektiv) eine bestimmte Menge von Kasusrollen, vergleichbar mit der syntaktischen Valenz. Eine Weiterentwicklung der Kasusgrammatik ist die Funktionale Grammatik von Dik [Dik91]. In dem Bereich der Dialogsysteme wird in vielen Fällen die *Slot-Filling* Technologie [Sou00] eingesetzt, die mit der Idee der Kasusgrammatik eng verbunden ist [Bru75, Min81]. Beispielsweise wird sie in dem *Air Travel Information Service Task* (ATIS) [Iss93] eingesetzt.

Seit einigen Jahren werden vermehrt Mikrofone in Dialogsystemen eingesetzt. Der Einsatz von Spracherkennern führt immer auch zu einem Unsicherheitsfaktor, da die Eingaben falsch erkannt werden können. Robustes Parsen ist in diesen Bereichen ein wichtiges Analysekonzept. Viele solcher Methoden kommen in telefonbasierten Auskunftsystemen oder zur Übersetzung von gesprochener Sprache zum Einsatz (z. B. in [Lav96, All96, Lav97, Wah00, War99, Dow93]).

Eine Variante ist, die Verarbeitung statt auf vollständige Sätze, auf Phrasen auszurichten (z. B. in [Blo00]) oder partielle Analyseergebnisse zu einem zu verbinden [Wor98a, Wor98b]. Andere Systeme nutzen direkt bei der Spracherkennung grammatikalische Informationen von Phrasen, um die Erkennungsrate zu verbessern und direkt an die Domäne anzupassen (z. B. [Bau01, Lan03]). Eine weitere Methode ist der Einsatz von statistischen Verfahren für die Sprachverarbeitung (u. a. in [Min96, Joh04]). Hier ist eine genaue Kenntnis des Korpus nötig, ebenso wird eine große Menge an Trainingsdaten benötigt. Daher sind diese Ansätze sehr zeitaufwendig und nur dann anwendbar, wenn die Dialoge gut umrissen werden können. Für schnelle nachträgliche Erweiterungen sind sie nicht geeignet. Um Spontansprache verarbeiten zu können, ist in einigen Systemen die verwendete Grammatik genau auf die zu erwartenden sprachlichen Phänomene ausgerichtet [Kro00, Pet99, McK98]. Konkrete Ansätze für die Verarbeitung von Sprache werden in Kapitel 4 ausführlicher beschrieben.

3.4. Fazit

In diesem Kapitel wurden wichtige Theorien und Formalismen vorgestellt, die in der Sprachverarbeitung zum Einsatz kommen. Sowohl die Repräsentationsformalismen als auch die Grammatiktheorien und Verarbeitungsmechanismen tragen entscheidend zum Gesamtkonzept der einzelnen sprachverarbeitenden Systeme bei. Auch wenn die einzelnen Theorien mit unterschiedlichen Gedankenmodellen entwickelt wurden, sind sie nicht immer vollständig voneinander abgrenzbar. Viele Ideen und Teilaspekte sind von anderen Theorien übernommen und integriert worden. Ebenfalls wurden die Konzepte mit der Zeit weiterentwickelt. Die CD-Theorie z. B. basiert zum einen auf einer Merkmalssemantik, nutzt aber gleichzeitig das Konzept der Netze. Ebenso werden in Grammatiktheorien Ideen übernommen, wie z. B. der Tiefenkasus der Kasusgrammatik in anderen Theorien.

Die Verarbeitungsstrukturen und die Repräsentationsmechanismen sind keine unabhängigen Konzepte, sondern stehen in Verbindung zueinander. Für eine komplexe Analyse wird in der Regel beides benötigt. Sowohl Verarbeitungsmechanismus als auch semantisches Konzept sind wesentliche Bestandteile des Verarbeitungssystems. Eine komplexe Analyse entsteht, wenn beide Bestandteile ineinander greifen. Dabei können nicht immer die Repräsentationsformalismen genau einem Verfahren zugeordnet werden. Dennoch ist es so, dass es Kombinationen gibt, die vorteilhaft für die Analyse sind. Beispielsweise war die Konzeptuelle Semantik von Jackendoff als ein Beschreibungsmittel für die generativen Semantiken gedacht, kann sie doch gerade in der Dependenziellen Grammatik sinnvolle Verwendung finden (vgl. [Lob93b, Lob93a]).

Hier wird deutlich, dass eine Theorie für sich allein genommen wenig Aussagekraft für den Einsatz in der Mensch-Maschine-Kommunikation besitzt. Durch die Verwendung in realen Systemen können Schwachstellen und Grenzen aufgedeckt werden, aber auch die hilfreichen Eigenschaften spiegeln sich dort wieder. Daher werden im nächsten Kapitel diese Konzepte und Theorien direkt unter der Prämisse der Verwendung in Dialogsystemen betrachtet. Dort wird ein vollständigeres Bild aufgezeigt, wie Repräsentationsformalismen konkret zum Einsatz kommen, wie semantische Beschreibungsmittel mit den Verarbeitungsmechanismen kombiniert werden können und wie diese zusammenspielen, um eine möglichst nutzbringende Sprachverarbeitung zu erzeugen.

4. Sprachverarbeitungssysteme

Bisher existieren nur wenige Beispiele von Robotersystemen mit komplexen sprachlichen Interaktionsfähigkeiten. Dagegen bilden „klassische“ Dialogsysteme in der Mensch-Maschine-Kommunikation ein breites Spektrum an Realisierungen ab. Diese Systeme wurden jedoch für sehr unterschiedliche Einsatzgebiete und Anwendungen entwickelt und sind weniger gut mit Systemen der Sprachverarbeitung für mobile Roboter vergleichbar. Um das hier entwickelte Verfahren dennoch mit anderen Verarbeitungsmechanismen vergleichen zu können und einen Überblick über den derzeitigen Stand der Entwicklung von Dialogsystemen mit Fokus auf Mensch-Roboter-Kommunikation zu geben, werden in diesem Kapitel einige prominente Systeme vorgestellt. Sie wurden exemplarisch nach Kriterien ausgewählt, bei denen ein oder mehrere Schwerpunkte der hier vorgestellten Arbeit auf deren Systemkonzept zutreffen. Dabei sind folgende Aspekte von zentraler Bedeutung: Verarbeitung von Spontansprache, Multimodalität, Sprachumfang und Korpus, Flexibilität, Situiertheit sowie Robustheit. Sie bilden die Kernpunkte der Sprachverarbeitung von mobilen Robotersystemen mit sozialen Fähigkeiten in realen Weltszenarien (siehe Kapitel 7). Die folgenden Darstellungen der Systeme konzentrieren sich auf die Umsetzung und die Realisierung des automatischen Sprachverstehens, das Hauptziel dieser Arbeit.

Als zwei der ältesten Systeme werden das System SHRDLU und der Roboter SHAKEY vorgestellt. Verbmobil ist ein unimodales System, das in einem groß angelegten Projekt mit dem Ziel entwickelt wurde, ein Sprachübersetzungssystem für spontansprachliche Dialoge in mobilen Situationen zu erstellen. Die Sprachverarbeitung des virtuellen Roboterarms CORA legt seinen Fokus auf die Aspekte der Situiertheit und der Verarbeitung von Spontansprache. Mobile Robotersysteme werden ausführlicher behandelt, da die Entwicklung einer Sprachverstehenskomponente für ein mobiles Robotersystem den Schwerpunkt der hier vorgestellten Arbeit bildet. Diese Systeme sind von den Rahmenbedingungen am ehesten mit der in dieser Arbeit entwickelten Verstehenskomponente vergleichbar.

4.1. SHRDLU

Eines der ersten sprachverarbeitenden Systeme ist SHRDLU [Win72]. Es beantwortet Fragen, führt Kommandos aus und nimmt neue Informationen in die Wissensbasis auf. Das System agiert in der so genannten *blocks world*. Die simulierte Welt besteht aus unterschiedlichen Bauklötzen, die sich auf einem Tisch befinden. Winograd geht davon aus, dass für die Modellierung von

Sprachverstehen verschiedene Aspekte zusammen behandelt werden müssen. Im Vordergrund steht nicht die Behandlung einzelner Aspekte wie Syntax, sondern die Konzeption eines Gesamtsystems, das über syntaktisches, semantisches, kontextuelles und physikalisches Wissen verfügt. SHRDLU besteht daher aus einer Reihe miteinander agierender Komponenten, wie z. B. einem Parsingsystem, einem Lexikon, einer Semantikkomponente und einer Planungskomponente. Die Semantikkomponente verbindet die Analyse und die Planung miteinander. Der Planer basiert auf dem Resolutionsprinzip und nutzt für die Analyse Theoreme über die Welt, wobei in Form von Theoremen physikalisches Wissen und das Wissen über das Erreichen von Zielen gespeichert wird. Anweisungen sowie Fragen werden durch den Parser unter Verwendung der Theoreme in Ausdrücke übersetzt, die von der Planung weiterverarbeitet werden können. Dabei stellen die Theoreme Prädikate dar, mit deren Hilfe die eingetippten Sätze zu einem Handlungsplan generiert werden können. Antworten auf Fragen werden mit Hilfe eines Generators für Äußerungen produziert.

Das System SHRDLU weist bereits komplexe Kommunikationsfähigkeiten auf, die für die frühe Entwicklungszeit herausragend waren. Es besitzt zwar nur einen geringen Wortschatz, verwendet geschriebene und grammatikalisch korrekte Sätze, bindet jedoch verschiedene sprachliche Ebenen in den gesamten Verarbeitungsprozess ein und integriert zudem physikalische Eigenschaften von Objekten aus dem Diskurs mit ein.

Ein ähnliches Verfahren nutzt auch der mobile Roboter SHAKEY [Nil84, Col69], der sich in speziell präparierten Räumen bewegen kann und im Raum befindliche Blöcke auf Anweisung verschieben kann. Auch dieses System basiert auf dem rein symbolisch orientierten Ansatz und verwendet sowohl Pattern-Matching als auch einen Theorem-Beweiser, um die eingegebenen Sätze zu verarbeiten.

4.2. Verbmobil

Verbmobil war ein groß angelegtes Projekt unter Beteiligung verschiedener Universitäten und Konzerne aus der Industrie. Dabei wurden viele Bereiche der automatischen Sprachverarbeitung bearbeitet, mit dem Ziel, ein Sprachübersetzungssystem für spontansprachliche Dialoge in mobilen Situationen zu erstellen. Neben den Szenarien der Reiseplanung und der „remote PC maintenance“ war das prominenteste Ziel, ein automatisches Übersetzungssystem für Terminabsprachen zu entwickeln. Die im Folgenden beschriebenen Aspekte von Verbmobil beschränken sich deshalb vorwiegend auf diesen Bereich.

In den Dialogen des Terminabsprachesystems werden Vereinbarungen eines geschäftlichen Termins bearbeitet. Die Basissprache für die beteiligten Personen ist Englisch, jedoch sollen die Beteiligten auch in ihrer jeweiligen Sprache (Deutsch oder Japanisch) kommunizieren können. Dazu wird eine spontansprachliche Äußerung erkannt, interpretiert und in die Zielsprache Englisch übersetzt. Als Szenario für das System wird angenommen, dass ein Deutscher und ein Japaner hauptsächlich in Englisch kommunizieren und bei Problemen sich in der jeweiligen Mutterspra-

che an Verbmobil wenden, so dass das System dann das Gesagte ins Englische übersetzt und somit den Fortgang des Dialoges sichern kann. Die Gesamtdarstellung des Projektes findet sich in [Wah00].

Die diversen Funktionalitäten von Verbmobil (z. B. Spracherkennung, syntaktisch-semantische Analyse, Synthese) sind in verschiedene autonome Module aufgeteilt, die jeweils geschlossene Teilaufgaben bearbeiten. Der Aufbau des Gesamtsystems entspricht einer nicht-hierarchischen Multiagentenarchitektur. Die Architekturübersicht in Abbildung 4.1 zeigt die wesentlichen Module und Schnittstellen des Verbmobil Forschungsprototypen.

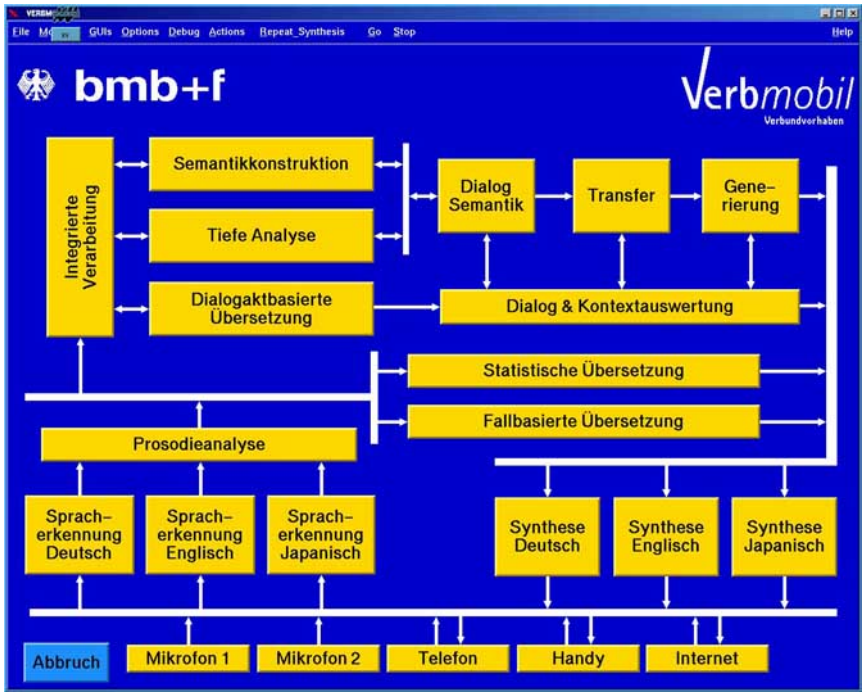


Abbildung 4.1.: Die Gesamtübersicht des Verbmobil-Projekt (aus [Wah00] S. 10).

Die Verarbeitung der Äußerungen durchläuft dabei einen mehrstufigen Prozessverlauf. Nachdem die Spracherkennung die Äußerung in einen Worthypothesengraphen (WHG) transformiert hat, werden zunächst mögliche Erkennungsfehler, Füllwörter und Versprecher heraus gefiltert. Anschließend werden drei Parser parallel für die Weiterverarbeitung der Äußerungen genutzt, die sowohl auf *flachen* als auch auf *tiefen* Analysestrategien basieren.

Für die flache Analyse fügt eine Komponente zum parallelen Parsing zunächst partielle syntaktisch-semantische Elemente in einen Interpretationsgraphen ein. Dabei werden die Äußerungen durch stochastische endliche Automaten in partiell syntaktisch und semantisch interpretierbare Einheiten aufgebrochen. Anschließend werden die Elemente durch Parsing-Strategien weiterverarbeitet. Der Chunk-Parser kann die Äußerungen am schnellsten verarbeiten, er stellt eine robuste Analyse bereit, liefert jedoch auch die am wenigsten akkuraten Ergebnisse. Da bei der Formulierung und Erkennung von Spontansprache Fehler gemacht werden und es häufig keinen

syntaktisch korrekten Pfad im WHG gibt, liefert diese Analyse in diesen Fällen mehrere einzelne Teil-Ergebnisse für die Äußerungen. Ein weiterer eingesetzter Parser ist ein probabilistischer LR-Parser, dessen Ergebnisse, wie auch die des Chunk-Parsers, durch eine semantische Analyse nachbearbeitet werden. Die syntaktisch-semantischen Analysemethoden sind lexikalisch gesteuert (*lexical driven*) und verwenden Regeln auf Basis der *Treebank* Formate ([Wah00] S. 200-215). Die sprach-abhängigen semantischen Datenbanken enthalten entsprechende Assoziationen zu einem Prädikatnamen, einer semantischen Klasse und einem Subkategorie-Frame. Parallel zu den flachen Parsing-Strategien wird ein HPSG-Parser eingesetzt. Dieser liefert die semantischen Strukturen automatisch mit, benötigt jedoch die meiste Rechenzeit und ist gegenüber spontansprachlichen Phänomenen und Spracherkennungsfehlern wenig robust. Ein Kontrollmechanismus des Auswahl-Moduls wählt aus den Ergebnissen der Parser mit Hilfe eines Bewertungsmechanismus das letztendlich beste Ergebnis aus.

Zusätzlich ist eine Dialogverarbeitung in Verbmobil eingebaut, um die Dialogakte ausfindig zu machen. Es nutzt ein Dialoggedächtnis und Domänenwissen zur Bereitstellung der Informationen. Ein Transfer-Modul bildet die ermittelte Bedeutung des Gesprochenen auf semantische Strukturen ins Englische ab, welche die Grundlage für die Generierung der Ausgabe bilden.

Mit dem hybriden Ansatz erreichte der Verbmobil-Forschungsprototyp 1.0 eine „approximativ korrekte“ Übersetzungsrate von 72,2% [Wah97]. In knapp drei von vier Übersetzungen ist der vom Sprecher intendierte Inhalt der Äußerung erkannt und in die Zielsprache übersetzt worden. Die Übersetzungsstrategien mit der flachen Analyse erreichten eine Erfolgsquote von 46,9 bzw. 46,3%. 51,9% der Übersetzungen konnte mit Hilfe der tiefen Verarbeitungsstrategie korrekt geleistet werden. Dabei lieferte in 12,3% nur die tiefe Verarbeitungsstrategie eine korrekte Übersetzung (nach der Definition von Wahlster [Wah97]) und in 13,1% der Äußerungen nur die flache Analyse ein approximativ korrektes Ergebnis. Das Ergebnis wurde auf einer Testmenge von etwa 20.000 Äußerungen erzielt, die jedoch keine dem Spracherkenner unbekannten Wörter enthielten. Durch die Ergebnisse wird deutlich, dass jedes Verfahren für sich genommen nur etwa die Hälfte aller Äußerungen korrekt verarbeiten kann und für sich alleine nur unzureichende Ergebnisse liefern würde. Erst durch die Kombination der unterschiedlichen Verfahren wird die Leistung in diesem Fall wesentlich verbessert.

Das Verbmobil-Projekt hat auf dem Gebiet der deutschen automatischen Sprachverarbeitung eine Leitfunktion. Viele Forschungsbereiche wurden in dem Projekt bearbeitet. Wie auch das Robotersystem BIRON (siehe Kap. 5) bekommt das System Verbmobil spontansprachliche Eingaben, jedoch unterscheiden sie sich darin, dass sie keinen Bezug auf eine reale Umgebung nehmen. Es nutzt flache und auch tiefe Analyseverfahren, um das System gegenüber möglichen Erkennungsfehlern und auch im Hinblick auf Spontansprache robust zu gestalten.

Leider finden sich nur vereinzelt Angaben zum Laufzeitverhalten einiger weniger Systemkomponenten. Jedoch läßt sich durch die Komplexität und die Vielzahl der Module vermuten, dass die Verarbeitungszeit für die Interaktion in Echtzeit mit einem mobilen Robotersystem mit begrenzter Rechen- und Speicherkapazität problematisch sein könnte.

Aufgrund der Erfahrungen mit den Experimenten zwischen Probanden und Robotersystem ist es wichtig, dass der verwendete Wortschatz und der Äußerungsumfang flexibel erweiterbar und

anpassbar ist (siehe Abschnitte 2.2 und 10). Viele Verhaltensweisen und spezifische Äußerungsvarianten der Interaktionspartner zeigen sich erst im laufenden Betrieb, wenn das System schon im Einsatz ist. Da jedes Modul in Verbmobil für sich jeweils unterschiedliche Datenbanken mit verschiedenen Inhalten und Repräsentationsformalismen verwendet, ist die Erweiterung der Datenbanken arbeits- und zeitintensiv und kann gegebenenfalls auch zu Inkonsistenzen führen. Daher wäre der Einsatz eines Systems wie Verbmobil in einen mobilen Roboter nur beschränkt sinnvoll. Ebenfalls ist es aufgrund der geringen Anzahl von vorhandenen Sprachdaten für mobile Roboterszenarien nicht möglich, statistische Verfahren einzusetzen. Der Aufwand für die Kollektion geeigneter Dialoge in ausreichender Anzahl wäre sehr zeitintensiv und aufgrund der Erfahrungen, dass viele Äußerungen nicht vorhersagbar sind, im Vorfeld nur teilweise sinnvoll.

Weitere Systeme, die ebenfalls rein akustische Daten verarbeiten, und daher ähnlichen Rahmenbedingungen unterworfen sind, sind Auskunftsysteme wie z. B. das telefonbasierte Wetterauskunftsystem JUPITER [Zue00] oder das Zugauskunftsystem EVAR [Mas94]. Das System TRAINS-96 [All96] unterstützt die Routenplanung in einer Transportdomäne.

4.3. Der virtuelle Roboter CORA

Das Forschungsgebiet der „virtuellen Agenten“ beschäftigt sich ebenfalls mit der Realisierung fortgeschrittener Mensch-Maschine-Schnittstellen. Exemplarisch wird hier das virtuelle Robotersystem CORA ausführlicher beschrieben, da es sowohl Situiertheit als auch spontansprachliche Aspekte in seiner Konzeption berücksichtigt.

CORA [Pet99, Mil97] ist ein simulierter Montageroboter, der natürlichsprachlich gesteuert werden kann. Die Welt, in der sich der Roboter befindet, sowie seine Bestandteile, sind in einer virtuellen Welt dargestellt. Bildinformationen über einen Ausschnitt der simulierten Welt erhält er durch ein Modul, das Bilddaten aus einer virtuellen Kamera an der Position des Greifarmes, simuliert. Ebenfalls liegen Daten von simulierten Tastsensoren an den Greiferinnenflächen sowie von telemetrischen Sensoren vor. Das Gesamtsystem besteht grob aus drei Teilen: der Simulation, der Steuerung und der Schnittstelle zum Benutzer (siehe Abbildung 4.2).

Die Aufgabe des Roboterarmes ist, aus Teilen eines baufix® Baukastensystems mit Hilfe natürlichsprachlicher Kommunikation mit einem Benutzer vollständige Aggregate zusammenzubauen. Der Benutzer bekommt dabei die gesamte Szene dargestellt, d. h. er hat Sicht auf den Roboter und auf die Objekte. Ebenfalls kann er die Perspektive des Roboters einsehen. Das gesamte System basiert auf einem agentenbasierten Architekturkonzept. Die Steuerungsarchitektur besteht aus drei Komponenten [Pet99]: Das Basissystem ist für die korrekte Ausführung von Aktionen verantwortlich. Es ist durch Sensoren und Aktuatoren in die simulierte Welt eingebettet. Die deliberative Komponente stellt Handlungswissen und Wissen über bereits durchgeführte Aktionen zur Verfügung. Die sprachverarbeitenden Komponenten stellen schließlich die sprachliche Schnittstelle zum Benutzer bereit, wobei die Kommunikation über Tastatur und Bildschirm erfolgt. Eine Darstellung der gesamten Steuerungsarchitektur ist in 4.3 abgebildet.

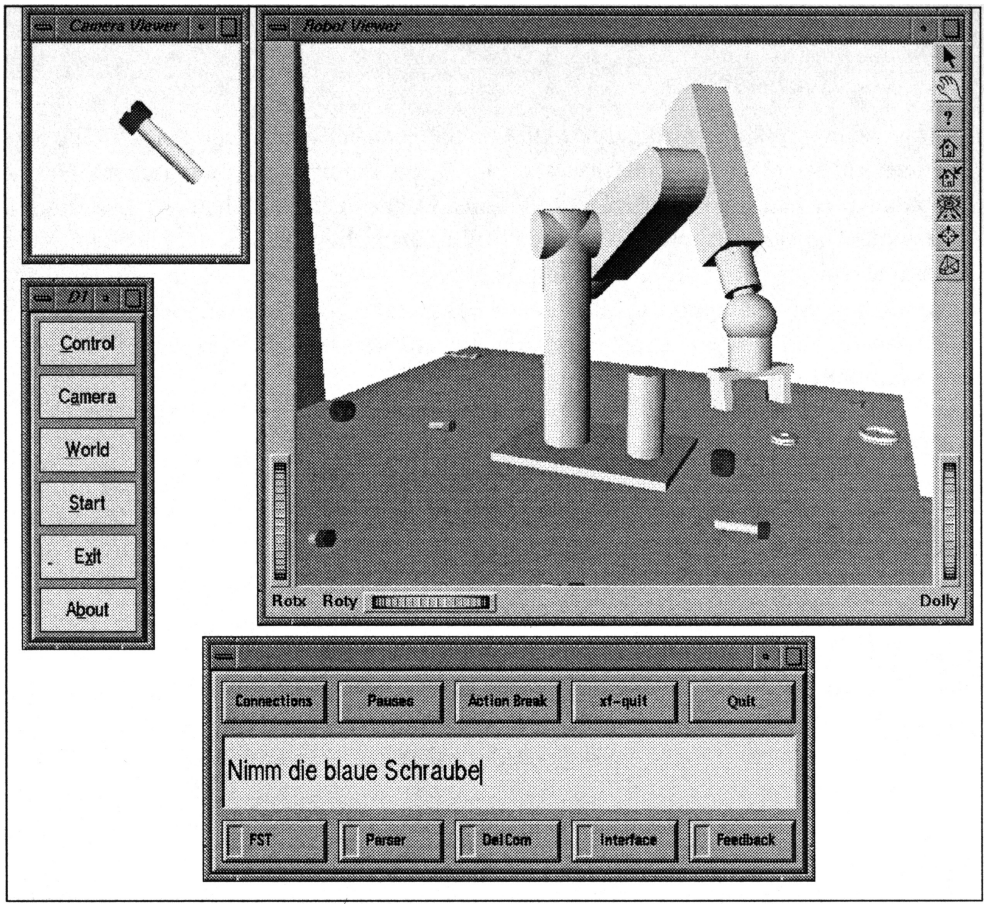


Abbildung 4.2.: Die Benutzeroberfläche des Systems CORA (aus [Pet99] S. 188).

Der Roboter besitzt in erster Linie *Ausführungskompetenz*, d. h. das Wissen und die Fähigkeit, die Objekte zu manipulieren und miteinander zu verbinden – es agiert entsprechend seiner Fähigkeiten teilautonom. Der Benutzer verfügt über das notwendige Konstruktionswissen und übernimmt die Aufgabe der Planung des Aggregates. Nur durch die gemeinsame kooperative Interaktion kann ein Aggregat konstruiert werden. Es ist daher von großer Bedeutung, eine gemeinsame Kommunikationsbasis zu schaffen.

Die Domäne des Systems ist handlungsorientierte Kommunikation in einem Konstruktionsszenario. Die Tatsache, dass der Mensch mit einem technischen System kommuniziert, beeinflusst ebenfalls die Art der zu erwartenden Äußerungen. Auch wenn die Existenz des ‘computer talk’ nicht gesichert ist, kann man Abweichungen von der Alltagssprache in gewissen Grenzen erwarten ([Pet99] S. 78). Die Anweisungen reichen von Einwortäußerungen bis zu komplexen parataktischen und hypotaktischen Konstruktionen. Viele Äußerungen sind nur unter Berücksichtigung der Gesamtsituation interpretierbar. Imperativsätze sind häufig zu erwarten, es ist jedoch auch mit indirekten Sprechakten zu rechnen. Ebenfalls müssen Nebensätze und Relativsätze verarbeitet werden können. Ellipsen treten aufgrund der Sprachökonomie der Mensch-Maschine-

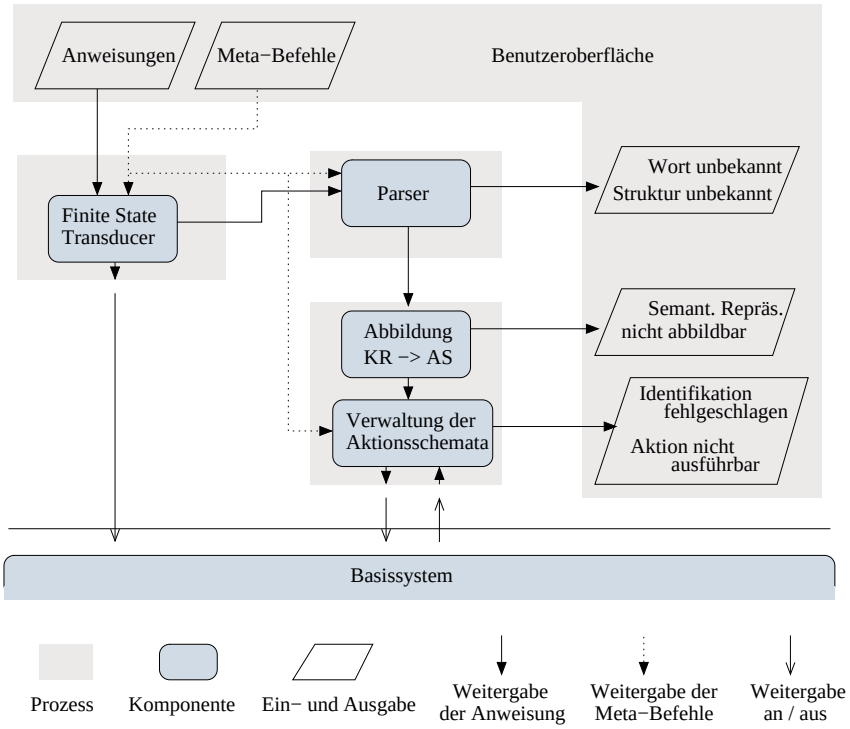


Abbildung 4.3.: Der Datenfluss bei der Verarbeitung von Anweisungen im System CORA (nach [Pet99] S. 192).

Kommunikation in diesem Kontext besonders häufig auf, z. B. wenn der Artikel weggelassen wird ([Pet99] S. 86). Sie spielen daher in dem System eine zentrale Rolle. Höflichkeit und andere Stilmittel, die den sozialen Kontakt ausdrücken, werden jedoch nicht erwartet. Die lexikalische Variationsbreite scheint ebenfalls nicht relevant zu sein, da sich die Benutzer auf Restriktionen im Vokabular eher einstellen können. Typische Beispiele für Äußerungen von Benutzern sind:

(1) „Drehe dich nach rechts!“

(2) „Die Schraube muss in die Leiste gesteckt werden.“

(3) „Nein, den anderen!“

Für die Verarbeitung von Sprache wird zwischen zwei unterschiedlichen Arten von Anweisungen unterschieden. Zum einen sind dies Anweisungen, die direkt in eine laufende Handlung eingreifen sollen und daher sofort zu verarbeiten sind. Diese Art von Anweisungen werden „Interventionen“ genannt und sind nur aus dem direkten Kontext verständlich. Beispiele hierfür sind: „Weiter nach vorne!“, „Nein, die andere!“ oder „Langsamer!“. Die andere Variante von Äußerungen werden als „Instruktionen“ bezeichnet. Sie erfordern vom System die Fähigkeit zur Handlungsplanung. Meist sind dies Anweisungen an das System wie z. B. „Verbinde die Leisten mit der blauen Schraube!“ oder „Lege den roten Würfel auf die Leiste!“. Sie werden vom System unterschiedlich genutzt: Während Interventionen in das Basissystem eingreifen, steuern Instruktionen die deliberative Komponente an.

Die Verarbeitung der unterschiedlichen Arten von Äußerungen erfolgt daher auf verschiedenen Wegen. Interventionen werden mittels eines *finite-state*-Transducers analysiert und die Ergebnisse direkt an das Basissystem übermittelt und dort verarbeitet. Die Interventionen weisen eine sehr einfache Struktur auf, die mit regulären Ausdrücken beschreibbar sind. Daher verwendet der Transducer hierfür ein Pattern-Matching-Verfahren.

Instruktionen sind all diejenigen Anweisungen, die nicht durch den Transducer erkannt werden können. Sie werden in eine semantische Repräsentation überführt, von der ausgehend sich Aktionsschemata spezifizieren lassen (vgl. [Lob98]). Dafür wird ein syntaktischer Parser eingesetzt, der neben der syntaktischen Struktur ebenfalls eine semantische Struktur aufbaut. Diese bildet die Grundlage für die Weiterverarbeitung im deliberativen System. Das zugehörige Lexikon enthält HPSG-nahe Einträge, die jeweils aus einem syntaktischen und einem semantischen Teil aufgebaut sind. Der unifikationsbasierte Parser beruht auf einem dependenzgrammatischen Ansatz [Goe96] (siehe Absatz 3.3). Bei der Analyse der Anweisungen werden die Abhängigkeitsbeziehungen zwischen den Wörtern betrachtet und überprüft. Wortstellungen werden mit diesem Ansatz nicht berücksichtigt, was eine große Wortstellungsfreiheit ermöglicht. Der Parser generiert aus den Instruktionen getypte Attribut-Wert-Paare, wobei die semantische Beschreibung parallel aufgebaut wird. Die Konzeption der Semantik ist dabei angelehnt an die konzeptuellen Strukturen im Sinne von Jackendoff [Jac90]. Nur dieser Teil wird an die deliberative Komponente weitergereicht, die die Langzeitziele verfolgt. Dabei werden die semantischen Informationen genutzt, um die korrespondierenden Aktionsschemata (siehe Kap.3.2.3) zu initiieren.

Folgende Handlungsmöglichkeiten bietet das Verhaltenssystem: *INSERT*, *PUT_DOWN*, *GRASP* sowie *MOVE* [Mil97]. Die Aktion ‘Legen’ (put) beispielsweise wird in die Basisaktionen *GRASP* und *PUT_DOWN* überführt. Ebenfalls werden die beteiligten Objekte den Aktionen entsprechend zugeordnet und die Informationen an die internen Sensoren weitergeleitet.

Beim Aufbau der semantischen Repräsentation werden nur die Informationen aus der Äußerung direkt verwendet, Weltwissen wird erst in der deliberativen Komponente eingesetzt. Aufgrund der Arbeitsweise der Dependenzgrammatik werden Wortstellungsinformationen nicht behandelt, jedoch werden Wörter, die nahe beieinander liegen bevorzugt, was für das Deutsche sinnvoll scheint. Im Lexikon sind etwa 800 Einträge als Vollformen sowie doppelte Einträge, die aufgrund des Parsingverfahrens notwendig disjunkte Attribut-Wert-Paarungen beinhalten. Dabei können einfache Deklarativ- und Imperativsätze sowie einige Formen von Nebensätzen verarbeitet werden. Grammatikalisch inkorrekte Äußerungen bezogen auf die Kongruenz sowie Äußerungen mit unbekannten Wörtern werden nicht analysiert.[Pet99]

In dem Szenario des Roboters CORA wird die gemeinsame Umwelt vom Roboter und Benutzer zwar nur simuliert, jedoch bieten sich auch hier einige interessante Ansatzpunkte für die Interaktion in einer realen Umgebung. Aspekte der situierten und spontansprachlichen Kommunikation spielen hier eine tragende Rolle. Dennoch bestehen in einigen Bereichen große Unterschiede, zum einen betrifft das die sprachlichen Fähigkeiten und zum anderen die generellen Rahmenbedingungen des Systems. Das Korpus von CORA wird zentral von Handlungsanweisungen für ein Konstruktionsszenario bestimmt. Sie können im Unterschied zur Kommunikation mit einem Roboter mit sozialen Fähigkeiten zum Teil komplexere syntaktische Strukturen aufweisen. Dage-

gen spielt Situiertheit in einem System, das in einer realen Umwelt interagiert, eine noch größere Rolle. Zeigegesten werden verwendet, das Umgebungswissen wird stärker in die Sprache einbezogen. Die Themen in der Kommunikation mit einem Service-Roboter bilden ein breiteres Spektrum ab, es werden mehr Freiheitsgrade in der Sprache erwartet und auch das Vokabular ist vermutlich größer. Es ist auch zu erwarten, dass sich gesprochene Sprache von getippter Sprache unterscheidet. In Kapitel 6 ist das Korpus von Dialogen zwischen Mensch und *Robot Companion* genauer beschrieben. Systeme in realen Umgebungen haben vermehrt mit Unsicherheiten zu tun, die Daten aus der Spracherkennung können fehlerbehaftet sein, ebenso die interpretierten Daten aus den visuellen Komponenten. Robustheit ist daher bei Systemen in einer realen Umgebung, im Gegensatz zu virtuellen Systemen, ein wichtiger Aspekt.

Neben CORA existieren eine Reihe weiterer virtueller Systeme, die über interaktive Fähigkeiten verfügen. Der virtuelle Agent MAX [Kop03, Kop05] ist ein antropomorpher künstlicher Agent, der einem Benutzer ebenfalls bei virtuellen Konstruktionsaufgaben assistieren kann und mit dem man *Smalltalk* halten kann. Eingaben werden mittels gesprochener Sprache [Kop03] oder per Tastatur [Kop05] eingegeben. Da auch hier ungrammatikalische Eingaben nicht ungewöhnlich sind, wird eine Methode der robusten Textanalye gewählt. Dabei werden im ersten Schritt generelle semantische Konzepte mittels Pattern-Matching herausgefiltert. Anschließend werden, ebenfalls mit Pattern-Matching und Schlüsselwortsuche, kommunikative Funktionen (bestimmte Wörter und Konzepte) herausgefiltert. Für die Verarbeitung der Sprache stehen 138 Interpretationsregeln zur Verfügung.

REA [Cas00] übernimmt die Rolle der virtuellen Immobilienmaklerin und interagiert mit einem Benutzer, um dessen Wünsche und Bedürfnisse zu erfassen. Sie zeigt ihm unterschiedliche Häuser und führt ihn darin herum, um letztendlich eine Immobilie zu verkaufen. Informationen werden über Tastatur eingegeben.

Die virtuelle Person Gandalf [Thó02, Thó99] kann Fragen über unser Sonnensystem beantworten. Dabei nimmt ein Datenhandschuh gleichzeitig die Gesten des Interaktionspartners auf. Ein grammatik-basierter Spracherkenner mit etwa 100 Wörtern und einer zusätzlichen Prosodieerkennung verarbeitet die Anfragen an das System. Sie werden mit Hilfe eines *template* basierten Mechanismus an die Wissensbasis des Systems weitergeleitet.

4.4. Mobile interaktive Robotersysteme

Im Gegensatz zu den bisher vorgestellten Systemen müssen mobile Robotersysteme mannigfaltige Probleme bewältigen, die Wissen aus vielen Disziplinen erfordern. Sie operieren – wie auch BIRON – in realen Welten und besitzen interaktive Fähigkeiten. Um die Gestaltungsvielfalt dieser Systeme zu veranschaulichen, wird im Folgenden ein breites Spektrum an Robotersystemen beschrieben. Dabei liegt der Schwerpunkt der Darstellung auf der Interaktionsfähigkeit und insbesondere auf der Verarbeitung von Sprache.

4.4.1. Der Roboter TJ

Der in den 90er Jahren entwickelte behaviourbasierte Roboter TJ [Con92] wurde von Torrance um ein Steuerungssystem und um eine Sprachverarbeitungskomponente erweitert [Tor94]. Er kann Wege in Gängen und Räumen eines Bürotrakts verfolgen und Zielpositionen ansteuern. Sein Gedächtnis umfasst eine Menge von Orten, und sog. *reactive-odometric plans* (ROPs), die die Orte durch Pfade miteinander verbinden. Insgesamt nutzt der Roboter diesen gerichteten Graphen, um einen Weg zu einem bestimmten Ziel zu finden.

Mit Hilfe der Sprache können neue Orte vermittelt werden, die der Roboter lernen und um die er seine innere Karte entsprechend erweitern kann. Die Sprachverarbeitung ist sehr einfach gehalten, die Aufforderungen müssen jeweils einem festen Muster folgen, die dann direkt in Lisp-Funktionen übersetzt werden. Dabei werden die zu verarbeitenden Äußerungen nicht durch gesprochene, sondern durch geschriebene Sprache eingegeben. Das System kann drei Arten von Sätzen verarbeiten:

- **Feststellungen** (*statements*), teilen dem Roboter die aktuelle Position mit. Mögliche Sätze sind z. B. „You are at Pat's office.“ oder „This is the conference room.“. Diese Informationen werden in den ROP-Graphen eingebaut.
- Aktionen führt der Roboter nach **Anweisungen** (*commands*) aus. Dabei werden im Wesentlichen die Geschwindigkeit, Richtung und das Ziel spezifiziert. Der Roboter kann Anweisungen wie „Turn left.“, „Go to the end of the hallway.“ oder „Go to Pat's office.“ verarbeiten.
- Zusätzlich kann der Roboter **Fragen** (*questions*) über seinen aktuellen Zustand beantworten, z. B. „Where are you?“.

Die Sätze, die TJ verstehen kann, sind in Form einfacher regulärer Ausdrücke definiert, die zusätzliche Variablen enthalten können (z. B. *place* und *direction*).

4.4.2. MOBSY

Der mobile Empfangsroboter MOBSY [Zob01] (siehe Abb. 4.4) wurde an der Universität Nürnberg-Erlangen für Besucher des Instituts entwickelt. Das Dialogsystem mit integriertem Sprachversteher basiert auf EVAR [Mas94] (siehe Abschnitt 3.2.2), das ursprünglich für die Fahrplanauskunft entwickelt wurde. Jedoch wird in MOBSY ein deutlich einfacheres System als in EVAR eingesetzt: Die Sprachverstehenseinheit sucht mit dem in [Nöt99] vorgestellten System in den erkannten Wörtern nach Phrasen, die für das System bedeutungstragend sind. Jede dieser Phrasen hat eine fest vorgegebene semantisch-pragmatische Repräsentation. Wörter, die für das System keine Bedeutung besitzen, werden ignoriert. Zusätzlich werden Pronomen nach einer einfachen Regel aufgelöst: Ein Pronomen wird durch das Objekt des vorangegangenen Satzes ersetzt, falls ein Satz im Dialogspeicher vorhanden ist. Ansonsten wird ein Fehler des Spracherkenners angenommen.

Abbildung 4.4.: Der Roboter MOBSY aus [Zob01].

4.4.3. JIJO-2

Der mobile Büroassistent JIJO-2 [Fry98, Mat99, Aso01] in Abbildung 4.5 ist für die Erledigung von Büroaufgaben konstruiert. Er soll Personen durch das Bürogebäude führen und Botendienste übernehmen. Um den Sprecher lokalisieren und Sprachsignale in einer Umgebung mit Hintergrundgeräuschen filtern zu können, besitzt der Roboter neben einer Reihe anderer Sensoren ein Mikrofon-Array. Das sprachverstehende System koppelt Spracherkennung und grammatikalische Analyse[1], indem es in den Spracherkenner für japanische Sprache einen in Lisp implementierten Left-Corner-Parser für kontextfreie Grammatiken zur syntaktischen Analyse integriert. Die Grammatik für die Büroumgebung enthält ca. 200 Wörter und 90 Erzeugungsregeln. Zusätzlich wurden zwei kleinere Grammatiken erstellt, zum einen für Antwortsätze, die Varianten der Zustimmung oder Ablehnung generieren kann, zum anderen eine Namens-Grammatik, mit deren Hilfe die Namenszuweisungen für Personen oder Orte erkannt werden. Die Eingaben der Interaktionspartner werden von drei Spracherkennern mit jeweils einer Grammatik und zugehörigem Vokabular verarbeitet. Die Dialogkomponente entscheidet abhängig vom Zustand des Dialoges, welche Art von Eingabe sie erwartet und wählt danach die Ausgabe des entsprechenden Erkennungsprozesses aus. Die Verarbeitung der Äußerungen geschieht annähernd in Echtzeit. Die semantische Analyse findet anhand eines Wortmodells statt. So ist für alle Verben und Adjektive im Lexikon verzeichnet, welche Argumente sie benötigen. Die Informationen werden an den Dialogmanager weitergereicht, um das Verhalten des Roboters zu steuern.

[1] wie auch frühere Implementationen der Sprachverarbeitung in BIRON [Top05]

Abbildung 4.5.: Der Roboter JIJO-2 aus [Mat99].

4.4.4. Das Projekt IBL

In dem Projekt IBL (Instruction-Based Learning) [Lau01] werden Lernmethoden eines Roboters erforscht, der sich in dynamisch veränderlichen Umgebungen befindet. Eines der Ziele ist, einem Roboter die Routenplanung mit Hilfe natürlichsprachlicher Eingabe zu lehren. Dafür wurden zunächst einmal Wizard-of-Oz Studien gemacht, um die Voraussetzungen der Interaktionsfähigkeiten zu schaffen. Ein Beispiel eines Wizard-of-Oz Szenario ist in Abbildung 4.6 dargestellt

Für die Erstellung des Korpus wurden Beispiele von aufgabenspezifischen Dialogen aufgenommen und analysiert und anschließend ein funktionales Vokabular generiert. Dieses Vokabular dient einerseits der Spracherkennung zur optimalen Erkennung, andererseits bildet es die Liste der primitiven Prozeduren, die die Benutzer in ihren Anweisungen verwenden. Eine Auswahl ist in Tabelle 4.1 abgebildet.

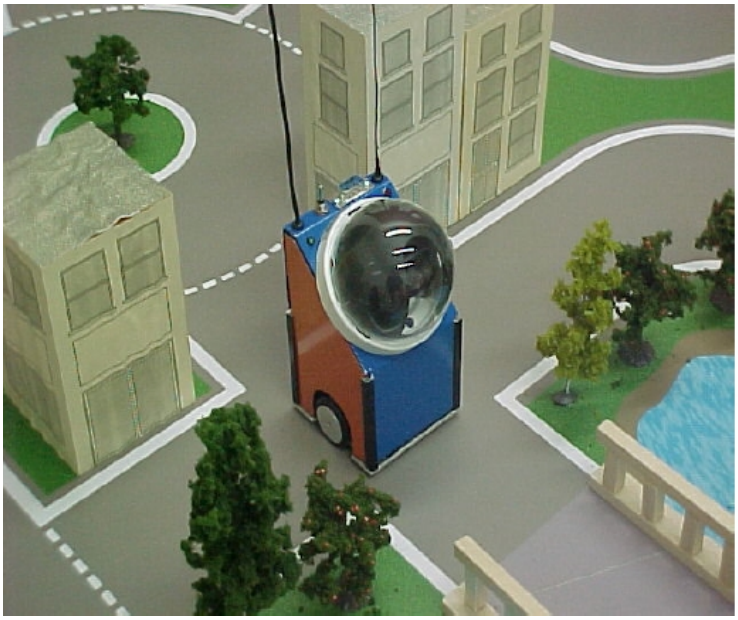


Abbildung 4.6.: Das Roboter-Szenario aus [Lau01].

	Anzahl	Primitive Prozeduren
1	308	MOVE FORWARD UNTIL [(past |over |across) <landmark>] | [(half_way_of | end_of) street] | [after <number><landmark> [left | right]] | [roadend]
2	183	TAKE THE [<number>] turn [(left | right)] | [(before | after | at) <landmark>]
3	147	<landmark> IS LOCATED [left | right |ahead] | [(at | next_to | left_of | right_of | in_front_of | past | behind | on | opposite | near) < landmark >] | [(half_way_of | end_of | beginning_of | across) street] | [between <landmark> and <landmark>] | [on <number> turning (left | right)]
4	62	GO (before | after | to) <landmark>
5	49	GO ROUND ROUNDABOUT [left | right] | [(after | before | at) <landmark>]
6	42	TAKE THE <number> EXIT [(before | after | at) <landmark>]
7	12	FOLLOW KNOWN ROUTE TO <landmark> UNTIL (before | after | at) <landmark>
8	4	TAKE ROADBEND [left | right]
9	4	STATIONARY TURN [left | right | around] | [at | from <landmark>]
10	2	CROSS ROAD
11	2	TAKETHEROAD in_front
12	2	GOROUND <landmark> TO [front | back | left_side | right_side]
13	1	PARKAT<location>
14	1	EXIT [car_park | park]

Tabelle 4.1.: Primitive Navigationsprozeduren erstellt aus dem IBL-Korpus aus [Lau01].

Ein Dialogmanager agiert als Interface zwischen Benutzer und Robotermanager. Er konvertiert Spracheingabe in semantische Repräsentationen [Lau02] und transformiert Anfragen des Robotermanagers zu Anfragen oder Antworten an den Benutzer. Für die Erkennung der Anweisungen wird der Spracherkenner Nuance eingesetzt, der auf Basis eines Grammatikmodells die erkannten Äußerungen direkt in eine logische Struktur umwandelt. Die Grammatik der Spracherkennung, erstellt mit Hilfe einer Spezifikationssprache (GSL), wird aus einer linguistisch motivierten Unifikationsgrammatik erzeugt [Bos02]. Sie kodiert das linguistische Wissen auf der Grundlage der Korpusanalyse und ist optimiert für die Domäne der Routenplanung für mobile Roboter. In dem IBL-Korpus sind einige hundert Wörter im Lexikon, wobei jedem Wort eine formale semantische Repräsentation zugeordnet ist. Durch diesen Aufbau arbeitet die Spracherkennung und das Parsing innerhalb eines Systems, die Spracherkennerergebnisse sind direkt logische Strukturen.

Das System interagiert nicht direkt mit den Benutzern, es bleibt daher die Frage offen, wie die Benutzer mit dem Roboter in der echten Interaktion reagieren und sich äußern würden.

4.4.5. Der Roboter CARL

Der mobile Roboter CARL [Lop03a, Lop05] ist der Prototyp eines intelligenten Service Roboters (siehe Abb. 4.7). Er besteht in seinem Grundelement aus einer *Pioneer PeopleBot* Hardware-Plattform der Firma ActiveMedia. Auf der mobilen Plattform ist ein Laptop mit einem Touch-Screen-Display montiert. Ebenfalls montiert ist ein Mikrofon für die Spracheingabe, eine Kamera zur Erkennung der Umgebung und ein Lautsprecher für die Sprachausgabe. Sprachinformationen können alternativ auch über eine virtuelle Tastatur eingegeben werden, die auf dem Touch-Screen-Display dargestellt ist. Ein animiertes Gesicht zeigt entsprechende Emotionen. Mit der Entwicklung der Roboterplattform CARL wurde nicht nur das Ziel eines mobilen autonomen Roboters verfolgt. Darüber hinaus sollte das System auch Aufgaben erledigen und lernen können. Sprachverstehen wird vorwiegend eingesetzt, um Fragen beantworten zu können („question answering“).

Die zentrale Steuerung des Systems ist „event-driven“. Je nachdem, welche Signale und Informationen eingehen (z. B. Sprache oder Bild), wechselt der Mechanismus in den entsprechenden Zustand. Dabei werden die Transitionsübergänge mit Hilfe von Prolog-Klauseln spezifiziert. Ein Inferenzsystem (induktives und deduktives Schließen), Sprachverarbeitung und Sprachausgabe sind in einem Modul implementiert. Ein weiteres Modul stellt Lernfähigkeiten bereit. Die Wissensrepräsentation für die Anfragen der Interaktionspartner basiert auf einem semantischen Netzwerk und auf Objekt Diagrammen in UML (Unified Modelling Languages). Mit den Netzwerken können Inferenzmechanismen leicht angewandt werden.

Das sprachverstehende System besteht aus verschiedenen Teilprozessen. Es verwendet eine Kombination aus *tiefer* und *flacher* Analyse, um die Robustheit des Systems zu erhöhen. Der Kommunikationsprozess ist modelliert als ein Austausch von Nachrichten, die Nachrichtentypen sind: *tell*, *ask*, *askf* und *achieve*. Zunächst konvertiert der Spracherkenner die Äußerungen in eine Folge von Worten. Anschließend wird mit Hilfe des LCFLEX [Ros98, Ros00] „left-corner“

Abbildung 4.7.: Der Roboter Carl aus [Lop03b].

Parsers basierend auf dem „Lexical Functional Grammar“ (LFG) Formalismus eine syntaktische Struktur erzeugt. Das Ziel ist, aus der Eingabe die vollständigste Struktur zu extrahieren und daraus eine Interpretation zu gewinnen. Dabei kann es auch Äußerungen interpretieren, die nicht vollständig durch die Grammatik abgedeckt sind, fehlende Teile werden dabei „übersprungen“. Kann das System keine sinnvolle Interpretation gewinnen, wird ein „memory-based learning“ Ansatz (MBL) gewählt. Er beruht auf der Annahme, dass intelligentes Verhalten aus analogem Schließen gewonnen werden kann. Dafür nutzt das System eine Menge von Beispielen als Eingabe. Daraus wird ein statistisches Modell gewonnen, das Eingabemuster klassifiziert. Anhand dieser Methode soll das System bedeutungstragende Informationen in Äußerungen identifizieren. Das System verwendet anstelle der direkten Wörter so genannte morpho-syntaktische „part-of-speech“ (POS) Einträge, die es mit Hilfe eines *Taggers* aus einer *Penn Treebank* erhält.

Um die Sprachverarbeitung mit dem Inferenzsystem zu verbinden, wandelt die semantische Analyse die syntaktischen Ergebnisse anschließend in eine prädikatenlogische Form um. Aus der Anfrage wird ein Ergebnis inferiert und eine Antwort für den Interlokutor generiert.

Dieser mobile Roboter kann in seinem Aufbau am ehesten mit dem Robotersystem BIRON (siehe Kap. 5) verglichen werden. Er enthält verschiedene Perzeptoren zum Wahrnehmen visueller und akustischer Informationen, und besitzt ein Spracherkennungssystem für Eingabe von Äußerungen. Das Robotersystem verwendet komplexe Mechanismen für die Sprachverarbeitung.

In gewissem Rahmen werden dabei auch spontansprachliche Phänomene und fehlerhafte Eingaben des Spracherkenners abgefangen. Beispielsweise sind Artikel in Äußerungen ein häufiges Problem bei der Spracherkennung, da sie nicht immer korrekt erkannt werden. Diese Problematik berücksichtigt das System und kann auch bei fehlerhaften Eingaben sinnvolle Antworten generieren.

Der Roboter agiert in einer realen Umgebung. Dennoch ist das System weniger auf situierte Dialoge ausgerichtet als auf die Aufgabe, Fragen zu beantworten („question answering“). Leider werden keine direkten Informationen über mögliche Diskursinhalte bereitgestellt. Indirekte Hinweise bieten die Beispiele von Anfragen oder Feststellungen an das Inferenz-System in [Lop05]: „What does Bob like?“ oder „Professor James is in France“. Aus den Beispielen kann entnommen werden, dass das System in den Dialogen weniger direkten Bezug zu seiner Umgebung nimmt und auch nicht über sie kommuniziert. Der sprachliche Kontext unterscheidet sich dadurch wesentlich von den Eingaben, mit denen der Roboter BIRON konfrontiert ist, da er sich direkt über die Informationen aus seiner Umwelt austauscht.

Fraglich am gesamten System ist der geringe Sprachumfang, der nur aus 36 Lexikoneinträgen besteht. Somit verfügt die sprachverarbeitende Komponente auch nur über eine sehr kleine und leicht wartbare Grammatik. Es ist daher schwierig, genaue Vorhersagen über das Laufzeitverhalten und den Aufwand der Erweiterungen für das System zu generieren, wenn der Umfang der möglichen Äußerungen deutlich größer wäre.

4.4.6. Weitere Robotersysteme

Neben den oben ausführlich beschriebenen Robotersystemen existieren eine Reihe weiterer Systeme mit Ausrichtung auf unterschiedliche Anwendungsszenarien. Sie werden hier in ihren Grundzügen vorgestellt. Da die Entwicklung mobiler Robotersysteme aufgrund der schwierigen Verarbeitung von Daten in realen Umgebungen hohen Designanforderungen gegenübersteht, handelt es sich meist um komplexe Systeme, die viele unterschiedliche Komponenten besitzen. Ein weites Spektrum an Herausforderungen sind bereits im Fokus der Roboterforschung, wie z. B. Personenaufmerksamkeit und -verfolgung, Objekterkennung und Routenplanung. Die Verarbeitung von Sprache ist jedoch bisher eher ein Randgebiet und wird nur von wenigen Forschergruppen in die Systemplanung einbezogen.

Viele Systeme sind auf spezielle Forschungsfragen fokussiert, die die Sprachverarbeitung nicht betreffen wie z. B. der Reinigungsroboter SINAS [vW01] oder die in [Bro90] beschriebenen Roboter. Die Forschergruppe, die u. a. den Roboter Leonardo [Bre04] entwickelt hat, arbeitet vorwiegend auf dem Gebiet der nonverbalen Kommunikation. Ebenso ist in [Sid03] das System Forschungsgegenstand der nonverbalen Interaktion, insbesondere der Bereich des *Engagements* zwischen Mensch und künstlichem System. Der Service-Roboter CERO [Hüt03, And99] ist das Ergebnis einer Design-Studie. Das Ziel ist hier, soziale und kollaborative Aspekte eines interaktiven Robotersystems unter Einbindung von potentiellen Benutzern zu untersuchen.

Schon die Spracherkennung ist aufgrund der Geräuschkulisse der Umgebung und der Eigengeräusche des Roboters ein komplexes Thema. Systeme, die sowohl Sprach- als auch Bildinformationen verarbeiten können, sind in ihrer Komplexität meist reduziert auf reine Spracherkennung wie z. B. in [Oku01, Tak98, Doi02].

Auch Systeme mit mehreren Sprachverarbeitungskomponenten sind oft eingeschränkt in Bezug auf ihren Sprachumfang wie z. B. der Interface-Roboter für Heimanwendungen Lino [Krö03, vB04] oder der Roboter HygeioRobot [Spi01], der ein Prototyp eines Couriers und Informationssystems in Krankenhäusern darstellt. Der Roboter Albert [Rog02, Ehr02](siehe Abb. 4.8), der Roboter CORA [Ios02] sowie der humanoide Roboter ARMAR [Asf00] aus dem Projekt Morpha [Lay01] sind Roboterassistenten, die gesprochene Kommandos verstehen können. Für diese Aufgabe wird ein Parser mit einer einfachen auf Kommandosprache optimierte Grammatik eingesetzt. Der mobile Roboter in [Fon03, Fon01] kann etwa 30 Kommandos verarbeiten, die nicht auf natürliche Sprache ausgelegt sind, sondern mittels eines PDA-Driver-Interfaces eingegeben werden.

Abbildung 4.8.: Der Roboter Albert aus [Ehr02].

Der mobile Service-Roboter Flo [Mon02, Roy00b, Roy00a] besitzt ein echtzeitfähiges Sprach-Interface. Für die Domäne der Kommando- und Kontrollaufgaben wird nur ein kleines Vokabular von 100 Wörtern benötigt, um die dem System gestellten Fragen beantworten zu können. Der Dialogmanager verwendet Keywordspotting über die Wortkette der erkannten Äußerung. In dieser Domäne ist komplexere Sprachverarbeitung nicht notwendig.

Robotersysteme für Umgebungen in denen sich viele Personen aufhalten, wie z. B. Museen, müssen stabil und robust konstruiert sein. Diese an sich beeindruckenden Robotersysteme besitzen keine direkten Möglichkeiten zu kommunizieren (z. B. RoboX [Tom02]) oder sind in ihren kommunikativen und interaktiven Fähigkeiten meist stark eingeschränkt. Beispielsweise wird der Roboter SAGE [Nou99] nur mittels Taster bedient. Der *Tour-Guide-Roboter* Minerva [Thr00, Bur99] kann seine internen Zustände nonverbal durch die Darstellung von Gesichtsausdrücken kommunizieren. Dagegen besitzt der autonome *Tour-Guide-Roboter* Jinny [Kim04] neben einem grafischen Interface ein Spracherkennungssystem, das mittels Keywordspotting Anfragen an den Roboter verarbeiten kann.

Es existieren einige Systeme für Konstruktionsszenarien, die im Gegensatz zu dem virtuellen Roboter CORA 4.3 zwar in einer realen Welt agieren, welche jedoch aus Komplexitätsgründen auf eine fest vorgegebene Umgebung reduziert wurde (z. B. Arbeitsplatte mit bekannten Objekten). Das in [Bau01] beschriebene System kann mehrere Bauteile zu einem Aggregat zusammenfügen. Die Welt, in der das System agiert, besteht aus realen baufix®-Bauteilen, die auf einem Tisch angeordnet sind. Hier wurde die Mobilität des Systems auf die Begrenzung der Tischplatte eingeschränkt. Für die Verarbeitung der Sprache wird ein Spracherkenner eingesetzt, der statistische Sprachmodelle mit einer LR(1)-Grammatik kombiniert und syntaktisch strukturierte Ergebnisse liefert [Wac98, BP99b]. Für das Verstehen der Anweisungen werden semantische Netze im Rahmen einer Erweiterung des Netzwerksystems ERNEST [BP99a] genutzt (siehe Kapitel 3.2.2). Ebenso handelt es sich bei dem Roboter KAMRO [Lue94] um einen Montageroboter, der um ein sprachverarbeitendes Front-End erweitert wurde (siehe Abb. 4.9). Die Hauptaufgabe der Spracheingabe ist die Steuerung der beiden manipulativen Arme. Diese können Objekte, die sich auf seiner Arbeitsplatte befinden, aufnehmen und ablegen. Die Äußerungen der Benutzer werden mit Hilfe eines unifikationsbasierten Parsers in Propositionen überführt.

Hermes [Bis02b, Bis02a, Bis99] ist ein Prototyp eines Service Roboters. Das Dialogsystem kann per Tastatureingabe oder natürlichsprachlich mittels gesprochener Sprache bedient werden. Ein Kommando-Interpreter verarbeitet die gesamte Eingabe. Er verwendet klassische linguistische Verfahren bestehend aus einem Parser, einer lexikalischen, einer syntaktischen sowie einer semantischen Analyse, die entsprechend hintereinander geschaltet sind. Eine geeignete Grammatik wurde für die speziellen Kommandosätze und Anfragen an den Roboter konstruiert. Beispielsätze, die der Roboter verarbeiten kann, sind „Turn around!“, „Grasp the ball!“ oder „Go to the kitchen!“. Ebenso sind Fragen, die ein bestimmtes Keyword (what, where, how) besitzen, erlaubt: „ What can you do?“ oder „Where are you?“. Nicht möglich sind aufgrund der eingeschränkten Syntax Äußerungen mit nachgestellten Information wie z. B. „Take the glass, the big one.“ oder „The glass over there, please take it.“ [Bis02b].

In dem WITAS Projekt [Lem01a, Lem01b] wird ein autonomer Helikopter mit Hilfe eines grafischen Interfaces oder per natürlicher Sprache gesteuert. Dabei kann das System Kommandos sowie Fragen verarbeiten. Für die Sprachverarbeitung wird wie auch im Projekt IBL [Lau01] der Spracherkenner Nuance eingesetzt, der gesprochene Sprache direkt in eine logische Form überträgt. Dafür wird eine Unifikationsgrammatik mit Hilfe des Gemini Compilers in eine kontextfreie Grammatik transformiert, die dem Spracherkenner als Basisinformation dient. Eine weitere

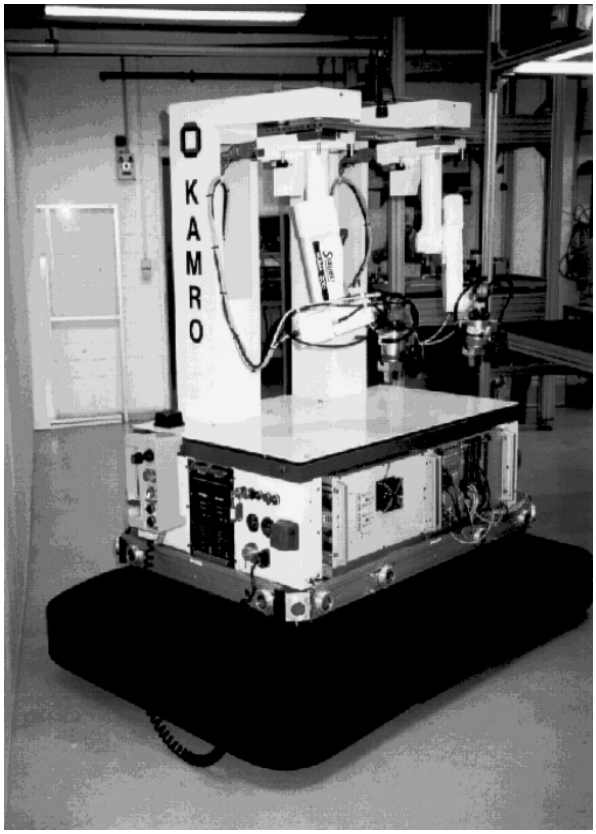


Abbildung 4.9.: Der Roboter KAMRO [Lae95].

Variante der Sprachverarbeitung im WITAS-Projekt wird in [Per01, Per99] beschrieben. Dort wird neben einem klassischen Spracherkenner ein syntaktischer Parser zur Verarbeitung von Kommandos eingesetzt. Die Ergebnisse der Sprachverarbeitung liegen in einer Lisp-ähnlichen Struktur vor. Insgesamt enthält das Vokabular des Systems 70 Wörter.

4.5. Fazit

Im ersten Teil dieses Kapitels wurden die Systeme beschrieben, die jeweils Einzelaspekte abdecken, die für die komplexe Mensch-Roboter-Kommunikation relevant sind. Aufgrund der sehr unterschiedlichen Zielsetzungen und der unterschiedlichen Rahmenbedingungen, für die die hier vorgestellten Systeme konzipiert wurden, ist es kaum möglich, direkte Schlüsse für die Aufgabenstellung der Mensch-Roboter-Kommunikation zu ziehen oder mehr noch, Verarbeitungsmechanismen ohne weiteres zu übernehmen. Es wird deutlich, dass das Design die genaue Problemstellung zu berücksichtigen hat. Dennoch werden einige Probleme in mono-modalen oder virtuellen Systemen behandelt, die ebenfalls in der Entwicklung des Sprachverstehens für kommunikative Robotersysteme berücksichtigt werden müssen.

Dem gegenüber stehen mobile Robotersysteme, die ebenfalls unter speziellen Aufgabenstellungen entwickelt wurden. Dabei spielen soziale Fähigkeiten und intuitive Kommunikation in den wenigsten Systemen eine zentrale Rolle. Dennoch sind mobile Roboter vielfältigen und komplexen Problemen gegenüber gestellt. Die Daten, die Roboter aus ihrer Umgebung aufnehmen, sind oftmals unsicher, es kann nie mit hundertprozentiger Wahrscheinlichkeit eine genaue Vorhersage getroffen werden. Informationen müssen gefiltert werden, um die Massen von Daten bewältigen zu können. Robotersysteme haben immer eine eingeschränkte Wahrnehmung, z. B. ist die Bilderrate pro Sekunde, die sie verarbeiten können, begrenzt oder die Farben verändern sich je nach Lichtverhältnis. Auch die Verarbeitung von Sprache ist problematisch - verschiedene Sprecher variieren im Farbklang, im Dialekt und im Frequenzspektrum. Je komplexer ein zu lösendes Problem ausfällt, um so schwieriger ist es, einen Roboter in Echtzeit reagieren zu lassen. Die Interaktion in realen Welten erfordert eine Vielzahl an Fähigkeiten, viele Komponenten interagieren miteinander und müssen aufeinander abgestimmt sein. Um die Komplexität der Aufgabe beherrschen zu können, besitzen die meisten Robotersysteme nur eine einfache Sprachverarbeitungskomponente mit stark eingeschränktem Wortschatz oder geringen Freiheitsgraden in der Äußerungswahl.

Die frühen Systeme SHAKEY und TJ können aufgrund ihrer geringen Rechenleistung nur wenige fest vorgegebene Sätze verstehen. In MOBSY wird eine Suche nach Schlüsselwörtern durchgeführt, die semantisch markiert sind, so dass sie direkt in eine semantische Darstellung überführt werden können. Diese Art von Systemklasse kann nur innerhalb einer sehr beschränkten Domäne sinnvoll eingesetzt werden. JIJO-2 verwendet eine syntaktische Analyse, die direkt an die Spracherkennung gekoppelt ist. Ein einfaches Valenzmodell stellt den Bezug zur Semantik her. Der Roboter CARL verwendet komplexe Mechanismen zur Sprachverarbeitung, besitzt jedoch nur einen geringen Wortschatz.

Keines der mobilen Robotersysteme ist jedoch auf einen großen Wortschatz ausgelegt und nur wenige darauf, dass die Interaktionspartner möglichst intuitiv mit dem Roboter kommunizieren können (z. B. durch Spontansprache wie in [Bau01]). Ebenfalls wurde keines der Systeme direkt im Hinblick auf seine Erweiterbarkeit konzipiert und auch die Entwicklung robuster Verarbeitungsmechanismen unter dem Aspekt der möglicherweise fehlerhaften Spracherkennung ist nur bei wenigen Systemen ein Thema (in [Bau01] und in dem WITAS-Projekt [Lem01b]).

Die Tabelle 4.2 zeigt einen Überblick über die hier vorgestellten Systeme, die über eine komplexere Sprachverarbeitung verfügen. Die einzelnen Systeme wurden nach Kriterien, die für die Entwicklung der Sprachverarbeitung von Robotersystemen mit sozialen Fähigkeiten wichtig sind, tabellarisch angeordnet.

Die Darstellung verdeutlicht, dass immer nur ein Teil der relevanten Bereiche von einzelnen Systemen abgedeckt ist. Es bleibt die Frage offen, wie ein sprachverstehendes System für den Bereich kommunikativer und sozialer Robotersysteme umgesetzt werden kann. Für den Roboter BIRON wurde daher eine eigene Sprachverstehenskomponente entwickelt, die alle zentralen Anforderungen berücksichtigt und weitestgehend umsetzt.

System	Eigenschaften					
	Mechanismus	**Sprachumfang**	**Spracheingabe**	**Spontansprache**	**Situierte Sprache**	**Echtzeitfähigkeit**
SHRDLU	synt. Parser	wenige reg. Ausdrücke	Tastatur	$-$	$-$	$-$
Verbmobil	komplexe Analyse	?	Mikrofon	$+$	$-$	?
CORA	komplexe Analyse	Handlungsanweisungen	Tastatur	$+$	$(+)$	$+$
TJ	Mustersuche	wenige reg. Ausdrücke	Tastatur	$-$	$-$	?
MOBSY	Mustersuche	reg. Ausdrücke	Mikrofon	$-$	$-$	$+$
JIJO-2	synt. Parser	200 Wörter, 90 Regeln	Mikrofon	$-$	$-$	$+$
Projekt IBL	synt. Parsing	ca. 200 Wörter	Mikrofon (indirekt)	$-$	$+/-$	?
CARL	LFG-Parser	36 Wörter	Mikrofon	$-$	$-$	$+$
Flo	Keywordspotting	100 Wörter	Mikrofon	$-$	$-$	$+$
System in [Bau01]	synt. Parser + sem. Netze	Handlungsanweisungen	Mikrofon	$+$	$+$	$+$
KAMRO	synt. Parser	?	Mikrofon	$-$	$(+)$	$+$
Hermes	komplexe Analyse	Kommandosätze	Tastatur o. Mikrofon	$-$	$(+)$	$+$
WITAS	synt. Parser	70 Wörter	Graf. Interface u. Mikrofon	$-$	$(+)$	$+$

Tabelle 4.2.: Vergleich der Systeme nach ausgewählten Kriterien

5. Das Robotersystem BIRON

Bei der Entwicklung des mobilen Roboters BIRON stand die Idee im Vordergrund, einen künstlichen Companion oder Partner mit sozialen und kommunikativen Fähigkeiten zu schaffen. Dieser Roboter soll sich unterhalten können, diverse Aufgaben übernehmen, sich neues Wissen aneignen, den Menschen in seiner Umgebung helfen und insgesamt möglichst universell einsetzbar sein. Ein entsprechender Roboter mit vielseitigen Fähigkeiten unterliegt einer besonderen Architektur. Viele verschiedene Systemkomponenten und Aufgabenbereiche fließen in das System ein und interagieren miteinander, um ein autonomes und homogenes Verhalten des Roboters zu erzeugen. Auch wenn bereits seit längerem humanoide Roboter erforscht werden und in Bezug auf die Mechatronik sehr leistungsfähige Systeme existieren (siehe z. B. [Has02, Hir98, Kan04, Lim00, Tev00]), stellt die Entwicklung eines sozialen und kommunikativen Roboters für die Interaktion zwischen Mensch und Robotersystem noch ein großes Problem dar. Die mobile Roboterplattform BIRON - der Bielefeld Robot Companion - [Lan03] geht wesentliche Aspekte in der Entwicklung eines künstlichen Partners an und bietet Lösungen für verschiedene Aufgabenbereiche, die innerhalb eines Robotersystems bestehen.

Ein offenes Architekturkonzept ermöglicht die Erweiterbarkeit, einerseits des Gesamtsystems um neue Aufgabengebiete und um zusätzliche Komponenten. Andererseits sind auch die einzelnen Module selbst erweiterbar. Die Sprachverarbeitung sollte z. B. leicht neue Sprachkonzepte und Wörter integrieren können. Für das Sprachverstehen selbst ist die Lösung für das Problem die Trennung zwischen Verarbeitungsmechanismus und Sprachdaten wie in Abschnitt 5.6.2 beschrieben.

Dieses Kapitel gibt einen Überblick über die Aufgaben und Fähigkeiten des Roboters BIRON, dessen einzelne Systemkomponenten und über die Gesamtarchitektur, die dieser Roboterplattform zugrunde liegt. Das sprachverstehende System, das in dem mobilen Roboter zum Einsatz kommt, wird hier nur kurz angerissen und in Kapitel 9 ausführlicher beschrieben.

5.1. Anwendungsszenario

Das Anwendungsszenario für den mobilen Roboter BIRON ist die so genannte *Hometour*. Die Idee dabei ist, dass der Roboter gerade erst in seiner neuen ihm noch unbekannten Umgebung angekommen ist. In der Anwendung soll der mobile *Robot Companion* durch die Interaktion mit dem Benutzer zunächst seine Umgebung kennen lernen und sich mit ihr vertraut machen. Dabei werden ihm die einzelnen Räume gezeigt sowie die Objekte in seiner Umgebung, die für

ihn in späteren Handlungen wichtig werden. Das Kennenlernen seines Umfeldes ist in etwa vergleichbar mit der Einweisung eines neuen Mitbewohners in eine Wohngemeinschaft. Auch dieser bekommt vermutlich erst einmal die Räume gezeigt, die Schränke, in denen sich gemeinsame Dinge befinden, aber auch Besonderheiten auf die er achten soll. Er erfährt z. B. Informationen, wo sich verschiedene Dinge befinden und auch Zusatzwissen, z. B. dass bestimmte Tassen nur für Tee zu benutzen sind oder eine Tasse besonders wertvoll für jemanden ist. Der Unterschied zum WG-Mitbewohner besteht darin, dass der Roboter zusätzlich noch ein eingeschränktes Weltwissen hat - er vielleicht gar nicht weiß, wie eine Tasse aussieht oder dass das Objekt vor ihm eine Gitarre ist. Darum ist die Einführung für den Roboter in seine Umgebung besonders wichtig. Erst im zweiten Schritt kann der Roboter dann den Personen seiner Umgebung helfen und ihnen Aufgaben abnehmen. Diese Aufgaben könnten z. B. das Blumengießen, Servieren von Getränken, Aufräumen oder auch die Unterhaltung der Mitbewohner sein, so wie sie in verschiedenen Robotersystemen angedacht sind [Aso01, Gra04, Dau04, Hüw04, Krö03, Nak01]. Als Beispiel dient hier ROBITA, die an Gesprächen über das Thema *Baseball* teilnehmen kann [Mat01]. Zudem besitzt ROBITA einen einfachen Arm, um auf Objekte in ihrer Umgebung zeigen zu können [Toj00].

Um seine Umgebung kennen zu lernen und um verschiedene Aufgaben erledigen zu können, benötigt der Roboter verschiedenste Fähigkeiten. Zum einen muss er in der Lage sein, Objekte zu lernen: sie zu erkennen und ihre Eigenschaften zu speichern. Er muss Informationen verschiedener Modalitäten zusammenbringen können, die Ansicht einer Tasse mit dem Wort „Tasse“ verbinden, ihre Farbe und Form sowohl visuell als auch symbolisch lernen und speichern können. Er muss Personen detektieren können, erkennen, ob sie mit ihm sprechen oder auf ein Objekt zeigen, und er muss sowohl hinter einer Person her fahren als auch sich frei im Raum bewegen können um selbständig zu einer neuen Position zu gelangen. Die kommunikativen und sozialen Fähigkeiten, die solch einem System abverlangt werden, sind enorm. Er benötigt Wissen über Dialogverhalten und Wissen über den speziellen Sprachgebrauch in solch einem Kontext. Dabei ist auch die Verarbeitung co-verbaler Informationen wesentlich, in der Erkennung sowie in der Produktion.

Die Interaktion mit dem Roboter BIRON kann man sich in etwa wie folgt vorstellen: Zu Beginn beobachtet der Roboter seine Umgebung. Sind Personen in der Nähe, richtet er seine Kamera abwechselnd auf die entsprechenden Personen aus und fokussiert auf sie. Will eine der Personen Kontakt mit BIRON aufnehmen, so liegt die Annahme zugrunde, dass sie sich BIRON zuwenden und ihn ansprechen wird. Hat BIRON erkannt, dass eine Person mit ihm interagieren will, begrüßt er die Person. Nach der Einführung in den Dialog kann der Roboter weitere Anweisungen entgegennehmen. Er kann auch erzählen, welche Fähigkeiten er besitzt oder seinen Namen nennen. Durch den Befehl „folge mir“ wird die Motor-Steuerung des Roboters aktiviert und BIRON fährt der Person hinterher bis zur gewünschten Stelle. Sagt der Interaktionspartner etwas wie „ich zeige dir etwas“ wechselt der Roboter in den entsprechenden Zeigemodus und kann daraufhin neue Objekte lernen oder sie wiedererkennen. Mögliche Dialogbeispiele werden in Abschnitt 6 und 10 genauer beschrieben. Abbildung 5.1 zeigt eine mögliche Situation während der Interaktion mit dem Roboter.

Abbildung 5.1.: Eine typische Interaktionsituation: BIRON folgt seinem Interaktionspartner.

5.2. Technische Ausstattung des Roboters

Der Roboter BIRON besteht in seinem Grundelement aus einer *Pioneer PeopleBot* Hardware-Plattform der Firma ActiveMedia. Sie ist mit grundlegenden Komponenten ausgestattet, die ihm Mobilität und Autonomie verleihen. Sie enthält zwei Antriebsmotoren und zugehörige Antriebsräder, Wegaufnehmer zur Positions- und Geschwindigkeitsberechnung und mehrere Sonar- und Anstoß-Sensoren zur Wahrnehmung von Hindernissen. Der Zugriff auf die Hardware-Kontrolle erfolgt mittels eines integrierten Micro-Controllers. Strom erhält der Roboter durch einen Satz Hochleistungsakkus. Zusätzlich ist der Roboter mit zwei im Roboter installierten Intel® Pentium®-III-Rechnern verbunden, zwei weitere Laptops können an den Seiten des Roboters angebracht werden. Eine Funk-Ethernet-Karte und ein Fast-Ethernet-Anschluss sind vorhanden, um sowohl einen drahtlosen Zugriff als auch einen Zugriff über LAN-Kabel auf das Robotersystem von außen zu ermöglichen. Zudem ist eine Soundkarte in dem System angeschlossen, die zur Ansteuerung von zwei Lautsprechern im Turm des Roboters dient. Unter Verwendung eines Sprachausgabemoduls kann damit der aktuelle Interaktionsstatus des Roboters gegenüber dem Interaktionspartner artikuliert werden (siehe Abbildung 5.2).

Um jedoch den Ansprüchen eines *Robot Companion* zu genügen, wurde die Grundausstattung des Systems um zusätzliche Hardware-Komponenten erweitert. Zum einen wurde ein Touch-Screen-Display für die Interaktion zwischen Mensch und Roboter installiert. Da aus Gewichtsgründen das System keine zusätzlichen Manipulatoren besitzt, können neben dem Erfassen von

Benutzereingaben auch Zeigegesten simuliert werden (z. B. durch Anzeigen des Objekts aus der Umgebung auf dem Display, auf das der Roboter referenzieren möchte). Zusätzlich kann ein Gesicht dargestellt werden, das dem Roboter mehr Persönlichkeit verleiht. Mit Hilfe von markanten Veränderungen im dargestellten Gesicht ist es möglich, die internen Zustände des Systems zu visualisieren.

Eine Pan-Tilt-Kamera wurde oben auf dem Roboter in einer Höhe von etwa 142 cm montiert, um Bilder von der oberen Körperhälfte des mit dem Roboter interagierenden Menschen aufzunehmen. Diese kann sich jeweils um $100°$ horizontal drehen und $25°$ vertikal heben und senken. Dadurch kann sie auf den Interaktionspartner ausgerichtet werden, um ihn zu verfolgen, aber auch, um Feedback bezüglich der Aufmerksamkeit des Roboters zu geben. Da die Kamera ein eingeschränktes Sichtfeld hat, wurde zusätzlich eine iSight-Kamera für die Erfassung von Gesten auf einer Höhe von etwa 95 cm installiert. Mit ihrer Hilfe kann ein Tiefenprofil der Umgebung erfasst und somit die Objekt- und Gestenerkennung erleichtert werden. Für die Spracherkennung werden zwei AKG-Grenzflächenmikrofone verwendet, wie sie auch in Freisprecheinrichtungen für Telefone eingesetzt werden. Diese sind direkt auf unterer Höhe des Touch-Screen-Displays angebracht. Ein Laser-Entfernungsmesser ist vorne auf unterer Beinhöhe montiert und liefert Messwerte bis zu 32 m Entfernung. Mit ihm werden Hindernisse und mögliche Personen detektiert.

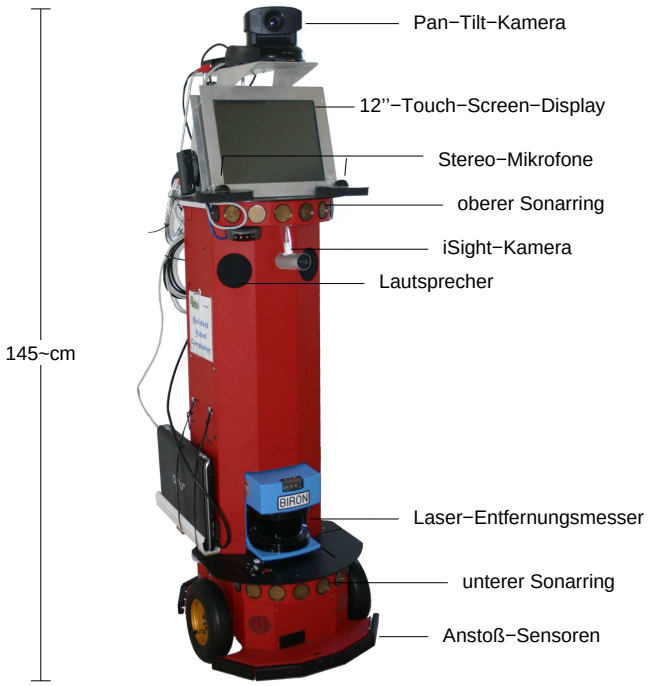


Abbildung 5.2.: Die Ausstattung des Roboters BIRON.

5.3. Kommunikationsframework der Roboterkomponenten

Aufgrund der hohen Komplexität des Roboters ist die Kommunikation zwischen den einzelnen Modulen ein wichtiger Aspekt. Viele Informationen fließen zusammen und werden untereinander ausgetauscht, um ein kohärentes Verhalten des Gesamtsystems zu erzeugen. Eine einheitliche Kommunikationsstruktur bietet ein gewisses Maß an Kontrolle über den Arbeitsablauf des Systems.

In der hier vorgestellten Roboterarchitektur wird das XML-basierte Kommunikationsverfahren XCF verwendet, das *XML-enabled Communication Framework* [Fri05, Wre04b], das den Austausch mittels der Beschreibungssprache XML unterstützt. Diese Beschreibungssprache eignet sich besonders als Basis für den Informationsaustausch, da sie sehr flexibel und zur Beschreibung von abstrakten Konzepten, aber auch von sehr unterschiedlichen Daten angemessen ist. Die XML-Sprache dient in unserem System zur Repräsentation von Zuständen, Ereignissen, Objekten und zur Vermittlung von sprachlichen Informationen. Daneben ist es mittels XCF möglich, zusätzliche Binärdaten wie z. B. Bilder zu übertragen, was die Repräsentation auf sub-symbolischer Ebene ermöglicht. Das System bietet diverse Kommunikationssemantiken, unter anderem *Publisher/Subscriber*-Mechanismen und synchrone und asynchrone Funktionsaufrufe. Alle XCF-Objekte und -Methoden können dynamisch zur Laufzeit im System registriert werden, was die Anbindung einzelner Module untereinander stark vereinfacht.

Die Ausgabe der Spracherkennung wird entgegen dem sonst üblichen XCF-Framework über das Kommunikationssystem DACS [Fin95] übertragen, das jedoch nicht auf die Übertragung von XML-Nachrichten ausgerichtet ist. Um dennoch das Gesamtsystem möglichst einheitlich zu gestalten, wurde eine zusätzliche Schnittstelle zwischen Spracherkennung und Sprachverstehen eingesetzt, die die Daten vom Spracherkenner in eine XML-Repräsentation konvertiert. Somit wird ein nachträglicher Austausch des Spracherkenners ermöglicht, bei dem nur die Schnittstelle geändert werden muss, die Verstehenskomponente selbst kann unverändert bleiben.

5.4. Das Szenemodell

Das Szenemodell dient der Speicherung von Informationen über Objekte, Personen und Handlungen aus der Umgebung, die dem Roboter bereits bekannt sind. Diese Informationen umfassen Attribute aus der Sprache wie Position oder Größe sowie visuelle Informationen, die durch die Objektdetektion geliefert werden. Das Szenemodell entspricht sozusagen dem Langzeitgedächtnis des Roboters, das zusätzlich noch Informationen unterschiedlicher Modalitäten zusammenführt. Psychologische Theorien der menschlichen Wissensrepräsentationen unterstützen die Annahme, dass symbolische Namen eines Objektes mit den sensorischen Eigenschaften wie z. B. der Bildansicht oder haptischen Eigenschaften [Bru66] verknüpft sind.

Zur Zeit ist das Szenemodell noch in der Entwicklungsphase und die benötigten Methoden und Funktionalitäten sind noch nicht vollständig umgesetzt. Anfragen an das gesammelte Wissen von BIRON oder auch das Abfragen von Objekten im Szenemodell sind in der aktuellen Implementierung noch nicht möglich, da der entsprechende Anfragemechanismus für das Szenemodell erst realisiert werden muss.

Das Szenemodell kann als eine multimodale Datenbank betrachtet werden. Zur Realisierung der Datenverwaltung ist die Verwendung eines *Active Memorys* [Wre04c] vorgesehen. Der Mechanismus beruht auf einer XML-Datenbank, auf die eine spezielle Server-Architektur aufgebaut ist. Damit ist es möglich, sowohl XML-Daten als auch binäre Daten zu verwalten und auf diese heterogenen Daten von verschiedenen Modulen des Robotersystems aus parallel zuzugreifen. Es ist wichtig, dass verschiedene Modalitäten eingebunden werden können. Beispielsweise beschreibt der Benutzer eine Farbe als Grün während das visuelle System auf Basis des HSV-Farbmodells arbeitet. Entsprechendes gilt für die Speicherung der Raumkoordinaten von Objekten. Diese Funktionalität ermöglicht es, Symbole in der real beobachteten Welt zu verankern, zu *grounden*.

Um unterschiedliche Formate ineinander überführen zu können, enthält das Szenemodell eine spezielle Komponente. Sie bedient sich einer so genannten Look-up-Tabelle (Nachschlagtabelle) mit drei Spalten: „Eigenschaft des Objektes“, „symbolische Beschreibung“ und „perzeptuelle Repräsentation“. Damit können beispielsweise Eigenschaften eines Objekts, wie „Farbe“ von der symbolischen Beschreibung wie „rot“ auf die sensorische Repräsentation wie z. B. die RBG- oder HSV-Werte abgebildet werden. Falls nun ein Modul eine Anfrage nach einem bestimmten Eintrag stellt, kann das System den entsprechenden Wert ausgeben. Somit ist eine Verknüpfung zwischen den Modalitäten möglich und das System kann überdies auch neue Beziehungen zwischen visuellen Repräsentationen und symbolischen Beschreibungen lernen.

Veraltete oder ungültige Informationen werden mit einem speziellen Mechanismus des *Active Memorys* automatisch entfernt und somit vom System „vergessen“. Dieser Mechanismus ist wichtig, da der Roboter in einer sich ständig verändernden Umwelt agiert und sich die interne Repräsentation so der Umgebung anpassen kann.

5.5. Die Aufmerksamkeitssteuerung

Damit der Roboter mit einem Menschen interagieren oder sich in einer realen Umgebung zurechtfinden kann, muss er die Personen und Objekte in einer Szene finden und in seinen Aufmerksamkeitsfokus bringen. Was für uns Menschen in der Regel ein Leichtes ist, stellt künstliche Systeme vor eine sehr schwierige Situation. Zunächst einmal müssen anwesende Personen detektiert werden und es muss erkannt werden, welche Person mit dem Roboter interagieren will und welche nicht. Zusätzlich müssen die im Raum vorhandenen Objekte detektiert werden, insbesondere die Objekte, über die gerade gesprochen wird - die aus Sicht von BIRON also im Fokus der Aufmerksamkeit sind. Diese beiden Fähigkeiten des Roboters werden im Folgenden vorgestellt.

5.5.1. Aufmerksamkeitssteuerung für Personen

Die für BIRON entwickelte Aufmerksamkeitssteuerung für Personen [Lan03] hat zum Ziel, einen geeigneten Interaktionspartner zu lokalisieren, der mit dem Roboter interagieren möchte. Zunächst wird jede Person, die sich dem System nähert, als potentieller Interaktionspartner angesehen. Sind mehrere Personen anwesend, muss erst der richtige Partner detektiert werden. Sprache kann dabei nicht als das alleinige Kriterium für den Kontakt angesehen werden, da sie auch einer anderen Person gelten kann und nicht unbedingt an den Roboter gerichtet ist. Daher werden multimodale Informationen wie Gesichtserkennung (z. B. die Richtung, in der die Person schaut), Beindetektion, Oberkörperdetektion und Sprecherlokalisation von der Personenverfolgung genutzt, um die Person herauszufinden, die mit der höchsten Wahrscheinlichkeit mit dem Roboter interagieren möchte (siehe Abb. 5.3). Wird auf diese Weise ein potentieller Partner gefunden, wechselt das System in den Modus der *Person Of Interest*. Dann wird nur noch diese Person als Interaktionspartner betrachtet und kann mit dem Roboter kommunizieren. Die Interaktion mit dem Roboter wird mit einer Begrüßung oder einer Anweisung initiiert und der Roboter wechselt in den Interaktionsmodus. Wendet sich diese Person wieder ab oder spricht nicht mehr, so kommen auch wieder andere Personen als Interaktionspartner in Frage. Ist keine Person im Fokus, richtet der Roboter abwechselnd seine Kamera auf die verschiedenen Personen im Raum aus. Mit dem Laser-Entfernungsmesser kann das System die Personen im Umkreis von 180° vor und neben BIRON detektieren.

Die Aufmerksamkeitssteuerung hat daneben auch die Aufgabe, den Personen in der Umgebung Rückmeldung zu geben, auf wen der Roboter fokussiert ist und wer als potentieller Interaktionspartner in Betracht kommt. Dazu dreht BIRON zuerst die oben befestigte Pan-Tilt-Kamera in Richtung der entsprechenden Person und im zweiten Schritt seinen Körper in Richtung der Person. Dadurch werden seine Sensoren so ausgerichtet, dass die Wahrnehmung des Systems möglichst optimal ist [Lan03, Haa04].

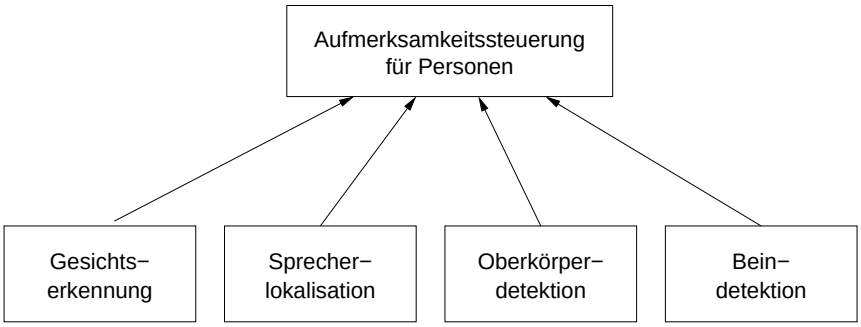


Abbildung 5.3.: Die Eingabe der Aufmerksamkeitssteuerung für Personen.

Zusätzlich zu der Ausrichtung der Kamera und der Roboterbasis wird von der Aufmerksamkeitssteuerung ein Gesicht auf dem Display angezeigt (siehe Abbildung 5.4). Das Gesicht verfügt über markante Merkmale wie Augen, Augenbrauen und Mund, um den internen Zustand des Roboters zu visualisieren. Diese haben für die menschliche Kommunikation eine große Bedeu-

tung und können beim Menschen das Verständnis für den internen Zustand des Systems erhöhen [Mae94, Eli02]. Durch das Ausrichten der Pupillen kann beispielsweise der Blickkontakt hergestellt werden oder auch angezeigt werden, dass der Roboter auf ein Objekt fokussiert. Geschlossene Augen signalisieren, dass der Roboter gerade im Ruhezustand ist und mit keiner Person in Kontakt ist. In diesem Zustand sind die rechenzeitintensiven bildverarbeitenden Module deaktiviert. Der Roboter kann durch Sprachperzepte daraus „geweckt“ werden und in eine allgemeine Bereitschaftsphase wechseln.

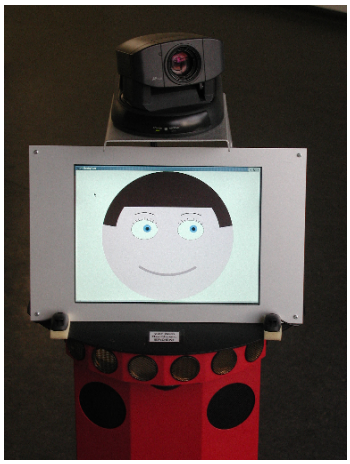


Abbildung 5.4.: Visualisierung der Aufmerksamkeit des Roboters über ein Gesicht, um interne Systemzustände in Form von Emotionen darstellen zu können.

5.5.2. Aufmerksamkeitssteuerung für Objekte

Gerade in dem Bereich der *Robot Companions* ist beispielsweise für das *Home-Tour-Szenario* eine möglichst intuitive Interaktion wichtig, ohne dass der Benutzer den Roboter extra trainieren muss. Ein weiterer Aspekt ist, dass der Roboter seine Umgebung kennen lernen, Veränderungen bemerken und in seine Wissensbasis integrieren kann. Die Aufmerksamkeitssteuerung für Objekte stellt dafür ein wichtiges Modul dar [Haa05]. Neben dem Finden von Interaktionspartnern ist es wichtig, Objekte und Orte herauszufinden, auf die der Benutzer in der Regel mittels verbaler Äußerungen oder deiktischer Gesten verweist. Für den Fall von BIRON ist das der Prozess, der es dem Roboter ermöglicht, auf das Objekt zu schauen, auf das der Benutzer verweist. Ist der Roboter in der Lage, auf Objekte zu fokussieren, neue Objekte zu lernen und bereits Gelerntes zu überprüfen und entsprechend zu aktualisieren, unterstützt das den Interaktionsprozess wesentlich.

Bei der Aufmerksamkeitssteuerung für Objekte wird zwischen bekannten und unbekannten Objekten unterschieden. Diese bedürfen verschiedener Verarbeitungsmechanismen. Semantisch bedeutet das die Trennung der Prozesse des Wiedererkennens und des Lernens. Besonders das Lernen ist eine schwierige Aufgabe, da die meisten Algorithmen ansichtsbasiert sind, was jedoch für unkontrollierbare Umweltbedingungen, wie z. B. in dem *Hometour*-Szenario, ungeeignet ist.

Daher wird der Schwerpunkt des BIRON-Ansatzes auf die Multimodalität gelegt - die Verwendung von Informationen über Gesten und Spracheingaben. Alle Informationen werden in der Wissensbasis, dem *Szenemodell* gespeichert.

Für das Erkennen und Lernen von Objekten wird zunächst die Gestenerkennung aktiviert [Haa05, Hof04]. Diese erkennt Bewegungen der Hand aufgrund der Trajektorie der Bewegungsrichtung. Ist das Objekt dem System bereits bekannt, wird zur Hand-Trajektorie auch die symbolische Darstellung des Objektes in die probabilistische Verarbeitung mit einbezogen. Wenn verschiedene Objekte in der Szene zu sehen sind, ist die Bewegungsrichtung der Hand zur Bestimmung des referenzierten Objektes wichtig. Die Gestenerkennung verwendet visuelle Eingabeinformationen auf Grundlage von Hautfarbensegmentierung und der Bewegungsanalyse. Daraus wird dann berechnet, ob der Benutzer eine deiktische Geste macht und auf welches Objekt referenziert wird.

Aufgrund der Positionsbestimmung wird nun die entsprechende Region des Kamerabildes ausgewählt. Bekommt das System vom Dialogmanager weitere Informationen wie die Farbe des Objektes, wird einerseits im *Szenemodell* geprüft, ob das Objekt bereits bekannt ist. Andererseits stellt das *Szenemodell* [Haa05] neben den bereits bekannten Eigenschaften und symbolischen Informationen auch entsprechende Bild-Muster des Objektes bereit. Bei bekannten Objekten wird ein einfacher Objekt-Erkenner [Lew95] verwendet. Ist das Objekt unbekannt, werden verschiedene Filter eingesetzt, um unterschiedliche Objekteigenschaften zu erkennen. Durch Zusatzinformationen aus der Sprache, beispielsweise über die Farbe, kann das entsprechende Objekt aus dem Bildausschnitt herausgesucht werden. Findet das System ein Objekt, wird eine Bestätigungsnachricht an den Dialogmanager geschickt. Werden mehrere oder keine Objekte erkannt, wird auch dies gesendet und der Dialogmanager stößt eine Anfrage an den Benutzer an, um weitere Informationen aus der Sprache über das Objekt zu erhalten (z. B. „das linke“ oder „das grüne“).

Die Aufmerksamkeitssteuerung für Objekte wird aktiviert, wenn der Benutzer auf ein Objekt referiert, z. B. mit einer Äußerung wie „das hier ist meine Tasse“, oder der Roboter selbständig mit einem Objekt interagiert. Dazu nutzt der Roboter die eingebauten Kameras: die iSight-Kamera, um die gesamte Szene zu überschauen und um die Position des Objektes relativ zum Roboter zu bestimmen, und die Pan-Tilt-Kamera für die genaue Detektion des Objektes. Neben den visuellen Eingabeinformationen (wesentliche Objekteigenschaften wie Farbe und Form) fließen weitere Informationen aus der Gestenerkennung, dem Dialogmanager sowie der Sprachverstehenskomponente ein. Das gesamte System nutzt dabei die Schlüsselinformationen aus dem Sprachverstehen, wie z. B. Hinweise auf mögliche Gesten, die durch deiktische Wörter wie „diese“ getriggert werden. Damit kann die Gestenerkennung explizit angestoßen werden. Die Koordination dieser Informationen sowie die Steuerung der Hardware (Kamera und Roboterbasis) wird mittels eines Endlichen Automaten realisiert. Der gesamte Mechanismus ist in [Haa05] detailliert beschrieben.

In diesem Zusammenhang wird deutlich, dass die sprachlichen Informationen einen wichtigen Beitrag auch für die Erkennung von Gesten und Objekten leisten können. Verbale Äußerungen sollten daher in solch einem Roboterszenario mit Rücksicht auf andere Modalitäten analysiert werden.

5.6. Sprachverarbeitung

Bei der Verarbeitung von Sprache auf der Roboterplattform BIRON sind hauptsächlich zwei Systeme beteiligt: zum einen die Spracherkennung, die die Audiodaten und Sprachperzepte erhält und in eine Wortkette transponiert und zum anderen das Sprachverstehen, das diese Wortkette verwendet und in eine semantische Interpretation umwandelt. Das sprachverstehende System ist stark von der Ausgabeperformance der Spracherkennung abhängig. Nur wenn eine bestimmte Prozentzahl an Äußerungen richtig erkannt wird, kann das sprachverstehende System darauf aufbauend korrekte Interpretationen erstellen. Jedoch sind aufgrund der erschwerten Rahmenbedingungen in diesem Szenario Erkennungsfehler kaum zu vermeiden (vgl. Kapitel 10). Daher wurde das sprachverstehende System auch mit der Absicht konzipiert, die Robustheit des gesamten Systems zu erhöhen. Wichtig ist, dass diese Spracherkennungsfehler erkannt werden, wenn möglich abgefangen werden und das Robotersystem auf potentielle Erkennungsfehler aufmerksam gemacht wird. Letztendlich hängt die Performance des gesamten Systems, das Erscheinungsbild insgesamt davon ab, wie gut die Sprachverarbeitung funktioniert und wie gut das System insgesamt mit auftretenden Problemen umgehen kann.

5.6.1. Spracherkennung

Der Benutzer soll eine möglichst intuitive Interaktion mit dem Roboter führen können. Daher sind anstelle von der Verwendung von Nahbesprechungsmikrofonen, wie beispielsweise bei der Nutzung eines Head-Sets, die Mikrofone direkt an dem Roboter montiert. Nachteilig für das System ist jedoch, dass dadurch der Signal-Rausch-Abstand geringer wird und es ist deutlich schwieriger, nur das Sprachsignal herauszufiltern und zu verarbeiten. Es bestehen daher größere Anforderungen an die Spracherkennungsleistung und generell an die Sprachverarbeitung als an Systeme, die das Sprachsignal direkt ohne Störquellen übermittelt bekommen. Das Phänomen, mit dem das System zurechtkommen muss, bezeichnet man als *Cocktailpartyeffekt*. Der Mensch ist in der Lage, auf einer Cocktailparty sich auf einen beliebigen Sprecher zu konzentrieren und ihn zu verstehen, obwohl alle durcheinander reden. Bisher war es nicht möglich, diese Perzeptionsleistung in ähnlicher Qualität auf ein Computersystem zu übertragen. Besonders problematisch bei der Spracherkennung sind folgende Aspekte:

- Da die Position des Sprechers variabel ist und der Abstand zum Mikrofon groß und nicht konstant, besitzt das Sprachsignal eine eher geringe Energie mit starken Schwankungen und einem ungünstigen Verhältnis von Signal- und Störkomponenten. Störgeräusche können ebenfalls mit aufgezeichnet werden und das Sprachsignal überdecken.
- Nicht jede Äußerung einer Person ist an den Roboter gerichtet. Da es keinen expliziten Knopf für das Ein- und Ausschalten der Aufzeichnung des Sprachsignales gibt, wie bei einigen anderen Systemen (z. B. [Bur98]), muss das System eigenständig erkennen, wann eine Äußerung innerhalb der Interaktion anfängt und wann das Ende einer Äußerung erreicht ist. Dabei können Fehler auftreten.

- Da das von der Spracherkennung verwendete Lexikon aufgrund der erschwerten Rahmenbedingungen auf das Aufgabengebiet des Roboters begrenzt ist, besteht die Gefahr, dass unbekannte Wörter oder ganze Äußerungen falsch oder nicht interpretiert werden können.

Diese Probleme sind typisch für Szenarien mit mobilen *Robot Companions* und müssen entsprechend berücksichtigt werden.

Für die Spracherkennung werden zwei handelsübliche Grenzflächenmikrofone eingesetzt (siehe Abschnitt 5.2), die auf BIRON montiert sind. Damit wird der gesamte vordere Bereich des Roboters abgedeckt. Für einzelne Testszenarien und in fremder Umgebung kann jedoch auch ein gängiges Head-Set-Mikrofon eingesetzt werden, um Störquellen zu reduzieren, wie beispielsweise auf dem IST-Event angewandt [Kle04].

Wendet sich der Benutzer zum Roboter hin und spricht ihn an, wird die Spracherkennung aktiviert. Die Kontrolle darüber liegt im Aufgabenbereich der Aufmerksamkeitssteuerung. Die situationsabhängige Aktivierung soll garantieren, dass nur Äußerungen, die an den Roboter gerichtet sind, verarbeitet werden. Dagegen werden Äußerungen, die einer anderen Person gelten, ignoriert. Die Erkennung der Blickrichtung bestimmt dabei, wann eine Äußerung dem Roboter gilt und wann der Benutzer zu einer anderen Person spricht.

Anschließend werden die ankommenden Sprachsignale zur Spracherkennung weitergeleitet und analysiert. Für die Spracherkennung wird ein inkrementelles sprecher-unabhängiges System eingesetzt, das auf Grundlage von Hidden-Markov-Modellen statistische und deklarative Sprachmodelle kombiniert [Wac98]. Um die Wortfolge zu finden, die mit maximaler Wahrscheinlichkeit produziert wurde, betrachtet man die Vorgänger des letzten Wortes und berechnet dann die wahrscheinlichste Wortkette. Dabei beschränkt man sich auf einen kleinen Teil der Vorgänger um der kombinatorischen Explosion bei der Bestimmung der Wahrscheinlichkeit eines Wortes P(w) zu begegnen. In der Regel wird, um gute Spracherkennungsergebnisse zu erhalten, bei einem solchen Verfahren ein großes Trainingskorpus eingesetzt. Für die Mensch-Roboter-Interaktion ist es jedoch nur mit einem sehr großem Aufwand möglich, solch ein Korpus zu erstellen. Es müsste eine Interaktion zwischen Mensch und Roboter simuliert werden, die einer realen Interaktion entspräche. Welche Kriterien dabei genau zu beachten sind, ist bis heute kaum erforscht. Daher wurde das akustische Modell für den Spracherkenner auf das „Wall-Street Journal“-Korpus (WSJ0) [Pau92] trainiert. Zusätzlich kommt ein Verfahren zum Einsatz, das die statistischen Modelle mit einer deklarativen LR(1)-Grammatik [Wac98] verbindet. Aufgrund der besonderen Problematik der Spontansprache wie in unserem Szenario kann keine festgelegte Grammatik definiert werden. Deshalb wird ein Parser verwendet, der neben ganzen Sätzen auch einzelne Konstituenten und auch einzelne Wörter akzeptiert.

Das Ende einer Äußerung wird mittels einer Stimm-Aktivitäts-Erkennung ermittelt. Da alle eingesetzten Systeme und Methoden fehleranfällig sind, können die Sprachdaten, die zum Sprachverstehen übermittelt werden, ebenfalls fehlerhaft sein oder wenn sie falsch abgeschnitten sind, auch nur partielle Äußerungen enthalten.

5.6.2. Sprachverstehen

Wie im vorherigen Abschnitt schon geschildert, kann eine korrekte Spracherkennung nicht garantiert werden. Das ist eine der Herausforderungen, der sich das Sprachverstehen stellen muss. Dabei stellt sich die Frage, wie automatisches Sprachverstehen mit diesen Parametern umgehen soll, um einen robusten Mechanismus zu erzeugen und somit den gesamten Dialogverlauf positiv zu beeinflussen. Desweiteren muss das System mit den spezifischen Anforderungen von situierter Spontansprache umgehen können und so weit wie möglich auch Hinweise auf Gesten des Interaktionspartners und Objekte in der Umgebung liefern. Auch ist es hilfreich, wenn eine Trennung zwischen dem Domänenwissen und dem Mechanismus selbst stattfindet.

Es wird daher bei BIRON bewusst ein System verwendet, das nicht auf einem syntaktischen Mechanismus beruht, sondern die semantischen Beziehungen innerhalb der Äußerungen in den Mittelpunkt stellt. Die semantische Kohärenz liefert außerdem einen guten Anhaltspunkt, um die Fehlerrate der Spracherkennung bestimmen zu können, also für die Qualität der Erkennungsleistung. In dem System dient das Wissen über semantische Beziehungen innerhalb verschiedener Situationen als Grundlage für die Verarbeitung von Äußerungen. Dafür wurden spezielle semantische Konzeptentitäten entwickelt, die so genannten *situierten semantischen Einheiten* („situated semantic units“ *SSUs*). Diese basieren auf dem Korpus situierter Kommunikation (siehe Kapitel 6) und sind auch in Hinblick auf deren automatische Verarbeitung konstruiert. In Kapitel 8 wird das Konzept der SSUs ausführlicher behandelt. Zusätzlich wurde ein spezieller Mechanismus erstellt, um diese Daten zu einer einzigen semantischen Einheit verarbeiten zu können.

Die Verarbeitung der Äußerungen basiert auf dem Grundgedanken der Theta-Rollen-Theorie (siehe Abschnitt 3.1.2). Dabei wecken einzelne Wörter Assoziationen zu bestimmten Konzepten. Ein „Ball“ besitzt z. B. verschiedene Eigenschaften, er ist rund und er hat eine bestimmte Farbe. Genauso entsteht bei dem Wort „rollen“ eine bestimmte Assoziation: Ein rollender Gegenstand, vielleicht eine Person, die Verursacher des Roll-Prozesses ist, eine Umgebung, in der etwas rollt. Das Wissen um diese semantischen Assoziationen wird genutzt, um innerhalb einer Äußerung Beziehungen zwischen den Wörtern und ihren Konzepten aufzubauen. Als Ergebnis der Verknüpfungen entsteht ein Netzwerk semantischer Entitäten.

Mit Hilfe der SSUs wird das Netzwerk von Relationen zwischen semantischen Konzepten in BIRON etabliert. Sowohl das Lexikon als auch die semantischen Konzepte sind in einer externen Datenbank gespeichert und werden erst während des Programmlaufs eingelesen. Die Roboterplattform selbst bleibt dadurch vom diskursspezifischem Wissen unberührt.

Bei der Verarbeitung wird die Wortkette nach den in Relation stehenden SSUs analysiert. Spezielle SSUs bilden dabei den Startpunkt des semantischen Parsens. Im Verlauf der Verarbeitung werden die SSUs miteinander verbunden um ein kohärentes Netzwerk zu schaffen. Je mehr Wörter miteinander verbunden werden können, umso besser wird die Spracherkennungsleistung bewertet. Letztendlich wird das Netzwerk als XML-Repräsentation an den Dialogmanager weitergereicht. Das genaue Verfahren wird in Kapitel 9 detaillierter beschrieben.

Dieser Mechanismus kommt dem Gesamtsystem gleich in mehreren Aspekten zugute: Zum einen ist es nicht auf grammatikalisch korrekte Äußerungen angewiesen. Dadurch werden den Benutzern viele Freiheitsgrade erlaubt, was gerade die Eigenheiten situierter Kommunikation unterstützt. Zum anderen liefert die semantische Kohärenz ein gutes Maß für die Bewertung der Ergebnisse aus der Spracherkennung. Auch kann das System die Interpretationen direkt an den Dialog weiterleiten, ohne aufwendig die Zwischenschritte und Berechnungen zwischen einer syntaktischen Analyse und einer semantischen Darstellung auszuführen. Desweiteren ist auch die Trennung von Weltwissen und Diskurswissen und dem eigentlichen Verarbeitungsmechanismus gegeben, so dass das System an sich unabhängig von der direkten Aufgabe ist und dieses Wissen problemlos austauschbar ist.

Mit dem Verfahren ist es auch möglich, die Assoziationen aus der Sprache zu anderen Modalitäten, z. B. der visuellen, herzustellen. Die Farbe „Rot“ kann mit einem RGB-Wert belegt werden oder das Wort „diese“ eine Stelle im Raum markieren, die mit Hilfe einer Geste angezeigt wird und somit eine Verknüpfung zwischen Umgebung und Sprache ermöglicht.

5.7. Die Dialogsteuerung

Der in dem Robotersystem integrierte Dialogmanager dient als Schnittstelle zwischen der Sprachverarbeitung und der Roboterkontrolle. Er steht auch in direkter Interaktion mit dem Benutzer: Er generiert die Antworten und bestimmt das Dialogverhalten. Dabei ist es das Ziel, Instruktionen von einem menschlichen Interaktionspartner zu erfassen und entsprechend zu reagieren. Die Architektur des Dialogmanagers ist agenten-basiert und erlaubt die Verarbeitung von multimodalen Informationen.

Die Sprachanalyse sendet die ermittelten Interpretationen einer Äußerung mit Rücksicht auf mögliche Zeigegesten oder Objekte im Raum an den Dialogmanager, die dann weiter verarbeitet werden. Übermittelt das Sprachverstehen mit der semantischen Interpretation beispielsweise Zusatzinformationen über eine potentielle co-verbale Geste, wird durch den Dialogmanager die Aufmerksamkeitssteuerung für Objekte explizit getriggert. Wenn das System dann eine Anweisung an den Roboter bekommt, z. B. dass der Roboter dem Benutzer folgen soll, wird diese Anfrage weiter an die Roboterkontrolle gesendet. Die Fehler der verschiedenen Komponenten laufen ebenfalls in diesem Modul ein, so dass der Benutzer auch über mögliche Probleme informiert wird. Wenn die Aufmerksamkeitssteuerung den Benutzer aus ihrem Fokus verloren hat, wird dies mitgeteilt und der Benutzer kann entsprechend reagieren, sich beispielsweise wieder dem Roboter zuwenden oder stehen bleiben.

Die Verarbeitung der eingehenden Daten basiert auf dem Begriff des *Groundings* [Li06a]. Damit ist der Austausch von Informationen zwischen beiden Dialog-Partnern auf der Grundlage einer gemeinsamen Basis gemeint. Die Kommunikationspartner signalisieren dem jeweils anderen, dass die Äußerung verstanden oder nicht verstanden wurde. Die Beiträge werden repräsentiert als „Interaction Units“ (Interaktions-Einheiten, kurz „IU“), die sowohl aus der verbalen als auch

aus anderen Modalitäten bestehen können. Ein Dialogabschnitt besteht immer als Paar-Beziehung zweier IUs, deren eine Seite die Rolle der Präsentation (*Presentation*), die andere die der Akzeptanz (*Aczeptance*) beinhaltet. Eine gemeinsame Basis (*Common Ground*) besteht, wenn solch eine akzeptierende IU existiert. Manchmal ist es notwendig, eine weitere IU zu erzeugen, beispielsweise durch eine Rückfrage, um im weiteren Schritt eine Akzeptanz-IU erzeugen zu können. Auf der Grundlage dieses Konzepts wird der Dialogfluss gesteuert. Ein Beispiel für eine IU ist die Frage nach dem Namen des Roboters. Nennt der Roboter dann seinen Namen, wird durch die Akzeptanz eine gemeinsame Basis aufgebaut. Wenn der Roboter sich ebenfalls nach dem Namen des Interaktionspartners erkundigt, wird eine neue IU erzeugt. Umgesetzt wird der Dialog mit Hilfe eines erweiterten *Push-Down Automaten*. Eine Bereitstellung entspricht dem Einfügen eines Eintrags in den Stack, eine Akzeptanz löscht einen Eintrag vom Stack.

Das System kann auf Anweisungen und Anfragen des Benutzers reagieren, gibt aber auch von sich aus Informationen an den Benutzer, die für wichtig erachtet werden (mixed-initiative). Die Ausgabe des Systems kann wiederum multimodal erfolgen, mittels eines verbalen und eines nicht-verbalen Generators. Dazu nutzt es die IUs als Informationsträger. Eine nicht-verbale IU des Roboters könnte nach der Äußerung „schau mal dort“ zusammen mit einer Geste das Bewegen der Kamera in Richtung der gezeigten Position sein. Für die Sprachausgabe verwendet das System Bausteine von Äußerungen, die in einer externen Datenbank gespeichert sind und je nach Zustand ausgewählt werden. Mit Hilfe der Sprachsynthesetools Festival und MBROLA werden die an den Benutzer gerichteten Äußerungen generiert.

Die Reaktion des Robotersystems selbst beeinflusst das Verhalten des Benutzers, der sich eigene Annahmen über die Fähigkeiten des Roboters macht. Ein sehr offensichtlicher Einfluss ist die Art der Sprachausgaben. Benutzer verwenden oft dieselben Wörter und Äußerungen wie das System. Das wiederum muss die gesamte Sprachverarbeitung berücksichtigen. Weitere Beobachtungen für den Einfluss des Verhaltens des Roboters auf den Benutzer werden in [Li04, Li06b] und in Kapitel 10 beschrieben.

Auf der Sprachseite nutzt das System die Informationen aus der Sprachverstehenskomponente anhand verschiedener Diskursinformationen. Es unterscheidet z. B. zwischen den Kategorien Anfrage, soziale Interaktion, Anweisung und Verneinung. Es kann dadurch sofort unterscheiden, ob der Benutzer beispielsweise eine Anweisung gegeben oder eine Frage gestellt hat und entsprechend reagieren. Für diesen Zweck wird die Theorie der Diskursstruktur nach Grosz und Sidner [Gro86] genutzt, um die Art der Äußerungen während der Interaktion zu beschreiben. Jedes Interaktions-Paar nach der Grounding Theorie entspricht einem Dialogakt [Li06a]. Demnach erhält der Dialogmanager von der Verstehenskomponente semantische Einheiten, die die einzelnen Diskurssegmente widerspiegeln und anhand derer es möglich ist, schnell und angemessen zu reagieren.

Es wird davon ausgegangen, dass beide Partner sich kooperativ verhalten, da beide ein gemeinsames Ziel verfolgen. Daher werden nur Dialogakte berücksichtigt, die diesem Anspruch entsprechen.

5.8. Dynamische Themendetektion

Um das Robotersystem BIRON im Hinblick auf Situationsbewusstsein zu verbessern, wurde ein dynamischer Themendetektor („Dynamic Topic Tracking“, kurz DTT) integriert [Maa06]. Erkennt das Robotersystem einen Themenwechsel, z. B. dass jetzt über die Küche gesprochen wird anstatt über das Wohnzimmer, kann er sowohl die Objekterkennung als auch die Spracherkennung unterstützen. Die Aufmerksamkeitssteuerung für Objekte kann damit beispielsweise die entsprechende Objektdatenbank auswählen und somit seinen Suchraum einschränken. Ist die Spracherkennung in der Lage, ihr Lexikon während des Betriebes auszutauschen, kann auch das durch den Themendetektor getriggert werden und so die Erkennungsrate verbessert werden.

Themenwechsel werden einerseits in der Sprache benannt, z. B. explizit durch Äußerungen wie „lass uns über die Küche reden“ oder implizit durch Anweisung zu einem Ortswechsel wie „komm mit“. Andererseits können Themenwechsel auch durch weitere Modalitäten markiert werden. Beispielsweise kann das Herumschauen einer Person einen Themenwechsel markieren - der Benutzer orientiert sich neu und redet dann vielleicht über ein neues Konzept oder einen anderen Gegenstand. Für die multimodale Themendetektion werden deswegen sowohl Informationen aus der Aufmerksamkeitssteuerung als auch Diskursinformationen genutzt, die das Sprachverstehen mit den semantischen Repräsentationen generiert (siehe Abschnitt 9).

5.9. Gesamtarchitektur

Die sinnvolle Integration aller Systemkomponenten in eine Gesamtarchitektur ist ein wesentlicher Aspekt bei der Erstellung eines flexiblen, multimodalen Systems, das die natürliche Interaktion überhaupt erst ermöglicht. Die Gesamtarchitektur von BIRON basiert auf einem hybriden Kontrollmechanismus bestehend aus drei Schichten: eine reaktive, eine intermediäre und eine deliberative Schicht [Fri05]. Abbildung 5.5 zeigt die Übersicht der einzelnen Schichten und deren zugehörige Module. Die Architektur ist so aufgebaut, dass weitere Module leicht in das System integriert werden können. Die Aufmerksamkeitssteuerung für Personen und die Aufmerksamkeitssteuerung für Objekte befindet sich in der reaktiven Schicht der Architektur.

Die Aufmerksamkeitssteuerung wird durch den Execution Supervisor konfiguriert, um verschiedene Verhaltensweisen anzunehmen. Dieser ist die zentrale Kontrollkomponente für den Roboter, in der mittels eines Endlichen Automaten der aktuelle Zustand des Roboters festgehalten ist. Bestimmte Ereignisse, z. B. Instruktionen vom Benutzer an den Roboter, werden dort abgebildet. Je nach Zustand werden dann die entsprechenden Verhaltensweisen des Roboters ausgelöst. Bei der Anweisung „folge mir“ beispielsweise geht der Roboter in den *Follow*-Modus über. Der Antriebsmotor wird angeschaltet und der Roboter versucht einen konstanten Abstand zum Benutzer beizubehalten, sprich, ihm zu folgen, ohne ihm zu nahe zu kommen oder ihn aus seinem Sichtfeld zu verlieren. Ein weiterer Zustand ist z. B. *Show*, bei dem der Roboter die Kamera etwas nach unten neigt, um deiktische Gesten und Objekte im Raum zu erkennen, die ihm der Interaktions-

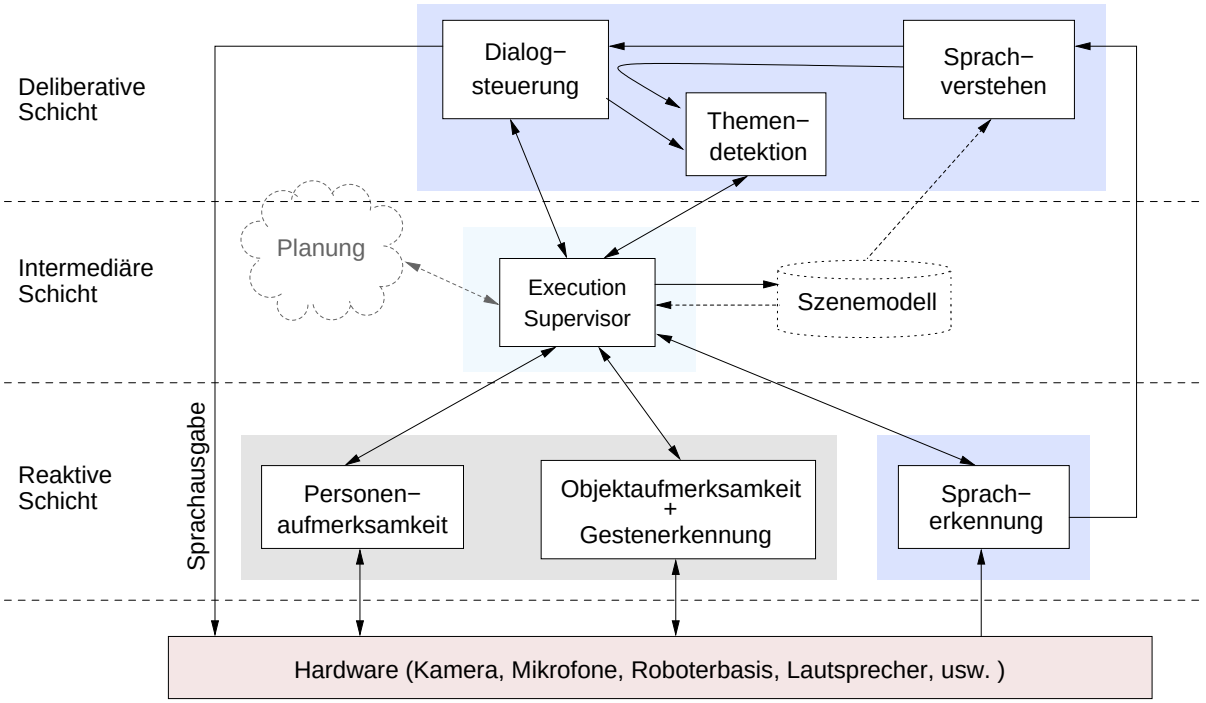


Abbildung 5.5.: Drei-Schichten-Architektur.

partner zeigt. Im *Person*-Modus richtet der Roboter die Kamera auf das Gesicht des Benutzers aus und wartet auf Anweisungen. Wenn der Dialogmanager signalisiert, dass auf ein Objekt in der Szene referiert wird, kann der Execution Supervisor die Kontrolle von der Aufmerksamkeitssteuerung für Personen zur Aufmerksamkeitssteuerung für Objekte verlagern, also den Fokus auf die interagierende Person verlassen und die Kamera auf das entsprechende Objekt richten. Dieses Verfahren schont die Rechenkapazität des Gesamtsystems, da damit spezielle rechenintensive Module für die Bildverarbeitung während der Laufzeit deaktiviert bzw. gezielt aktiviert werden können. Auch kann durch das Wissen über den internen Zustand des Systems getriggert werden, ob die Aufmerksamkeitssteuerung einen speziellen Interaktionspartner fokussieren soll oder alle Personen ihrer Umgebung abwechselnd betrachten soll.

Der Execution Supervisor speichert auch Daten zwischen, die die verschiedenen Module senden, um sie zu einem späteren Zeitpunkt an die entsprechenden Module weiterleiten zu können. Die eintreffenden Ereignisse werden anhand ihres zugehörigen Zeitstempels, der angibt, wann etwas erzeugt wurde, in eine Ereignis-Warteschlange einsortiert und dann chronologisch weiterverarbeitet. Da der Rechenaufwand bei der Informationsverarbeitung für verschiedenen Module gerade was die unterschiedlichen Modalitäten betrifft stark voneinander abweicht, können der Zeitstempel und die Ereignis-Warteschlange diese Unterschiede abfangen. Ebenfalls auf der Intermediären Schicht der Roboterplattform befindet sich das Szenemodell, das die multimodalen Daten für das System speichert.

Die Module der deliberativen Schicht sind die Dialogsteuerung, die Themenerkennung und die Sprachverstehenskomponente. Ein Planer für die Aufgaben der Navigation ist ebenfalls für diese Schicht vorgesehen. Zusätzlich soll die Planung von autonomen Handlungen dort angesiedelt werden. Die Dialogsteuerung erhält Eingaben von der Sprachverstehenskomponente und sendet gültige Instruktionen weiter an den Execution Supervisor. Der Themenerkenner greift ebenfalls auf diese Informationen aus dem Sprachverstehen zu und nutzt sie, um mögliche Themenwechsel zu erkennen.

5.10. Fazit

Die Qualität der kommunikativen und sozialen Fähigkeiten des Robotersystems hängt von den einzelnen Modulen, aber vor allem vom Zusammenwirken aller Bestandteile ab, das maßgeblich das homogene Gesamtbild des Roboters prägt. Durch die enge Verknüpfung der einzelnen Komponenten untereinander hängt die Leistungsfähigkeit eines einzelnen Moduls nicht nur von den direkt verbundenen Komponenten ab, sondern auch von indirekt beteiligten.

In einer realen Umgebung muss immer davon ausgegangen werden, dass das System verrauschte oder artefaktbehaftete Daten erhalten kann. Das Robotersystem muss auf diese Problematik eingehen. Beispielsweise hängt die Funktionalität des Spracherkenners von der Qualität der Aufmerksamkeitssteuerung ab. Nur wenn ein Gesicht erkannt wird, wird die Spracherkennung getriggert und kann Äußerungen verarbeiten. Bei dem sprachverstehenden Modul wurde zwar die Möglichkeit mitberücksichtigt, dass die Spracherkennung fehlerhaft sein kann, jedoch bedingt auch die Güte der Erkennungsleistung die Qualität der Interpretationen. Je mehr Äußerungen korrekt erkannt werden, umso besser ist das Ergebnis des Sprachverstehens (siehe Kap. 10). Das Dialogsystem wiederum benötigt die semantischen Strukturen, um einen erfolgreichen Dialog zu gestalten. Es ist abhängig von der Qualität der Daten. Es nutzt weiterhin die Bewertung des Sprachverstehens und reagiert entsprechend mit einer Antwort oder einer Rückfrage bei einer schlechten Bewertung oder bei fehlenden Informationen. Die Ausgabe des Dialogs wiederum beeinflusst das Verhalten der Interaktionspartner (z. B. die Wortwahl und Lautstärke) und somit die Eingabe für den Spracherkenner. Es existiert folglich ein Kreislauf, zu dem jedes Modul mit seinen Funktionalitäten beiträgt und auf die Verhaltensweise der anderen Komponenten einwirkt.

Der Themendetektor benötigt ebenfalls die Interpretationen aus dem Sprachverstehen, er nutzt sowohl die Wörter aus den Äußerungen als auch die Dialogakte, die das Sprachverstehen bereitstellt.

Das sprachverstehende System kann nur seine Fähigkeiten unter Beweis stellen, wenn auch die anderen Module leistungsfähig sind. Es ist direkt von der Spracherkennung und dem Dialogsystem abhängig, indirekt aber auch von allen anderen Modulen im Robotersystem, da ihre Fähigkeiten wiederum vom Spracherkenner und vom Dialogsystem benötigt werden.

6. Korpus und Domäne

Die Motivation für die Entwicklung des sprachverstehenden Systems ist, eine möglichst intuitive Kommunikation zwischen Mensch und Roboter zu ermöglichen. Maßgebliches Vorbild ist die Idee eines Roboter-Gefährten, der im Haus oder Büro assistieren und Aufgaben übernehmen kann. Dieser Roboter kann Kaffee kochen, Blumen gießen, Besucher im Haus herumführen, einen verlegten Schlüssel suchen, Briefe in die Büros bringen und noch vieles mehr. Zuerst muss er sich jedoch mit der Umgebung vertraut machen, bevor er anschließend selber Aufgaben übernehmen kann. Abbildung 6.1 zeigt eine typische Interaktions-Situation zwischen Mensch und Roboter. Dafür dient der Roboter BIRON als Forschungsplattform.

Abbildung 6.1.: Interaktion mit dem Roboter BIRON.

Bei der Entwicklung stellte sich zuerst die Frage, was genau die Besonderheit der Mensch-Roboter-Interaktion ausmacht. Um Näheres über das Korpus und den Sprachumfang zu erfahren, wurden verschiedene Wissensquellen genutzt. Zum einen wurden theoretische Überlegungen und das Wissen aus anderen Kontexten herangezogen, zum anderen wurden Studien und Experimente durchgeführt, sowohl im Vorfeld als auch in der direkten Interaktion mit dem mobilen Roboter BIRON.

In dem hier vorliegenden Kapitel werden die Dialoge beschrieben, die sich aus einem solchen Szenario entwickeln können. Die Datensammlung und Analyse stellt einen sehr wichtigen Bereich der Systementwicklung [Pot99] dar. Auf der Grundlage der Dialoge wird das Weltwissen

und das sprachspezifische Wissen abgebildet, das der Roboter benötigt, um seine Aufgaben erfolgreich erledigen zu können. Letztendlich bilden die Analysen der Dialoge die Grundlage der Entwicklung des automatischen Sprachverarbeitungssystems des Robotersystems. Zunächst sind die generellen Rahmenbedingungen des Systems zu berücksichtigen, z. B. in welchen Räumen sich solch ein Robotersystem aufhält, mit welchen Objekten es zu tun hat, wie ein typischer Dialog zwischen Mensch und Roboter aussehen kann und welche sozialen Kommunikationsmittel benötigt werden. Die Analyse erfolgt auch im Hinblick auf die Art der Dialogakte, die während eines solchen Dialoges stattfinden. Da sich das Korpus generell auf die Interaktion mit einem *Robot Companion* bezieht und sich nicht nur auf Handlungsanweisungen beschränkt, kann mit diesen pragmatischen Zusatzinformationen der Dialogmanager zusätzlich unterstützt werden. Ein weiteres Ziel ist, die Wissensdomäne so allgemein wie möglich zu gestalten, um einfache und schnelle Erweiterungen und Änderungen des Sprachumfangs zu ermöglichen. Die Art der Dialoge beeinflusst zudem das Systemdesign der Sprachverarbeitung im Generellen. Beispielsweise gibt es Rahmenbedingungen, wie robust der Verarbeitungsmechanismus im Hinblick auf syntaktische Eigenschaften der zu verarbeitenden Äußerungen sein muss und auch auf die zu erwartende Komplexität.

Da sich Roboter und Mensch in einer gemeinsamen Umgebung befinden und die Personen ohne Einschränkungen kommunizieren können, wird davon ausgegangen, dass die Menschen alle ihnen zur Verfügung stehenden Kommunikationskanäle einsetzen und ebenfalls die Umgebung mit in ihre Interaktion einbeziehen. Sowohl situierte als auch multimodale und spontansprachliche Kommunikation sind in dem Szenario wichtige Aspekte (vgl. Kap. 2.3). Beispielsweise ist robuste Verarbeitung von Spontansprache sowie die Berücksichtigung von Gesten und anderen Informationen aus der Umgebung relevant für eine erfolgreiche Interaktion. Wie die Studie in [Hüt03] zeigt, bevorzugen die meisten Probanden gesprochene Sprache (82%), gefolgt vom Touch-Screen (63%), jedoch auch Gesten (51%) und Kommandosprache (45%) sind gewünschte Wege, mit dem Roboter zu kommunizieren. Für eine möglichst intuitive Kommunikation sollten daher diese Interaktionswege bereitgestellt werden.

Es gibt nur wenige Studien, die sich mit der Interaktion zwischen Mensch und mobilem Roboter beschäftigen. In [Hüt03] wird eine Langzeitstudie über die Interaktion mit einem Büro-Roboter vorgestellt. Jedoch wurde hier nicht der Gebrauch von verschiedenen Modalitäten ausgewertet. Allerdings wurde berichtet, dass die Versuchspersonen dazu tendierten, Gesten und Sprache gemeinsam zu verwenden, obwohl ihnen gesagt wurde, dass Gesten vom System nicht erkannt werden können. In [Mil97] liegt der Schwerpunkt auf der Entwicklung eines situierten virtuellen Robotersystems - auch hier wurden spontansprachliche Phänomene im Kontext eines Konstruktionsszenario beobachtet.

Um detaillierte Informationen über die verwendeten Äußerungen und die Kombination von Sprache, Gesten und Referenzen zu Objekten in der Umgebung zu erhalten, wurde eine “erste” Benutzer-Studie durchgeführt, die in Abschnitt 6.1 beschrieben wird. Anhand des Experimentes wurde das Korpus situierter Dialoge gebildet. Der Umfang des Sprachschatzes wurde ergänzt durch eigene Annahmen über Dialoge zwischen Mensch und Roboter im *Hometour*-Kontext, die in Abschnitt 6.2 erläutert werden. Abschließend wurde das Korpus erweitert durch Experimente

über die freie Interaktion von Benutzern mit dem Roboter BIRON (siehe Abschnitt 6.3). Anhand der Analyse dieser Daten sowie der in der Literatur zitierten Beobachtungen wurden die Anforderungen für das Design des Sprachverstehens-Systems aufgestellt (vgl. Kap 7).

6.1. Korpus für das Deutsche: das Blumengieß-Szenario

Um genauere Informationen über den Sprachstil, den Satzaufbau und Wortschatz zu erhalten, den Benutzer bei der Interaktion mit einem mobilen Roboter verwenden, haben wir eine erste Benutzerstudie erstellt. Sie dient als erste Grundlage für die Wissensdatenbank und für die Modellierung der automatischen Sprachverarbeitung. In der Studie wurde ein Szenario nachgestellt, bei dem ein Roboter Anweisungen zum Blumengießen bekommt. Insgesamt wurden in der Studie 14 Dialoge von deutschen Muttersprachlern aufgenommen und analysiert. Die genauen Instruktionen sind in Anhang A.1.1 festgehalten. Da der Schwerpunkt unseres Systems auf der Situiertheit der Dialoge liegt und um eine möglichst freie Interaktion zu erlauben, wurde der Roboter durch einen Menschen ersetzt, der vorgibt, ein Roboter zu sein. Um die Probanden möglichst wenig mit vorgegebenen Äußerungen zu lenken und ein konsistentes Bild des Roboters zu liefern, beschränkte sich der „gespielte Roboter“ auf eine möglichst geringe Variation an Antworten (hauptsächlich Zustimmung durch „OK“ oder Rückfragen durch „Welche?“). Die Analyse des Korpus ist ebenfalls in [Hüw04] beschrieben. Es ist natürlich klar, dass eine Studie mit einem simulierten Roboter nicht alle Aspekte echter Mensch-Roboter-Kommunikation erfassen kann. Es ist möglich, dass Menschen in realen Roboterszenarien anders reagieren. Dennoch ist diese Studie ein wichtiges Instrument, um ein möglichst breites Spektrum von Interaktionen zu erfassen. Zudem ist sie notwendig, um ein erstes Modell erstellen zu können und nicht nur aufgrund theoretischer Überlegungen ein System zu konstruieren. Nicht zuletzt werden neben den lexikalischen und syntaktischen Informationen auch Informationen über das benötigte Dialog- und Weltwissen in diesem Szenario geliefert.

Die Aufnahmen zeigten eine große Bandbreite von charakteristischen Kommunikationsstrategien innerhalb des Szenarios. Dabei ist auffällig, dass die Dialogstruktur typischerweise in drei Abschnitte unterteilt ist: Die thematische Einleitung enthält eine Begrüßung sowie Hintergrundinformation. Im mittleren Abschnitt wird der Roboter über seine Aufgaben instruiert. Abschließend erhält der Roboter Zusatzinformationen und eine Verabschiedung. Ein typisches Dialogbeispiel ist in den Abbildungen 6.2 und 6.3 dargestellt. Daraus lassen sich erste Annahmen über die Dialogakte erstellen, die in dem Kontext benötigt werden. Soziale Rahmenhandlungen bilden einen Teil der Dialoge. Ein weiterer Bereich sind die Instruktionen an den Roboter und die Objektbeschreibungen. Zusätzlich können weitere Informationen während des Dialogs formuliert werden (meist zu Beginn), die für die Rahmenhandlung wichtig sind. Der Abschluss des Dialogs wird in der Regel ebenfalls explizit kommentiert.

Person: hallo, äh ich fahr in nächster Zeit in Urlaub für eine Woche –
und du sollst die Blumen gießen
Roboter: WELCHE BLUMEN?
Person: <diese Blume – in dem blauen Topf > –
einmal am Mittwoch Abend – zirka zehn Milliliter
Roboter: OK
Person: <und diese Blume>
Roboter: WELCHE?
Person: <ich zeige auf diese Blume >
Roboter: OK
⋮
Person: Das wärs

Abbildung 6.2.: Dialog-Beispiel: Typischer Ablauf

Person: OK – ich habe hier in dem Raum drei Blumen
ich zeige sie dir nacheinander – das ist die erste Blume –
die muss dreimal in der Woche gegossen werden – gut.
Roboter: OK
Person: die darf ruhig ein bischen trocken werden, aber ...
Roboter: WIE TROCKEN?
Person: der Boden darf einmal richtig trocken sein, aber dann –
muss er auch wieder gegossen werden.
Roboter: OK
Person: und dann haben wir noch die – Blume dahinter
Roboter: WELCHE BLUME?
Person: ja, die mit dem großen weißen Topf – genau – ...
die kann einen ordentlichen Schuss Wasser bekommen ...
⋮
Person: ja – das wärs

Abbildung 6.3.: Dialog-Beispiel: Umgangssprachlicher Dialog

Wie bereits von anderen Wissenschaftlern beobachtet [Blo95, Nig95, Kro00, Mil97], gibt es wesentliche Unterschiede zwischen gesprochener und geschriebener Sprache (siehe auch Kap. 2.4). Die Experimente bestätigten ebenfalls, dass Spontansprache oft nicht der Standardgrammatik und der Struktur geschriebener Sprache entspricht. Vielmehr wurde bei den Untersuchungen die speziellen Eigenschaften von situierter Spontansprache in dem multimodalen Roboterszenario deutlich. Die Probanden verwenden oft kurze Phrasen oder einzelne Worte. Pausen sind häufig inmitten einer Äußerung vorhanden. Sie sind in den Darstellungen der Dialogausschnitte mit einem „–“ markiert. Hesitationen, Abbrüche oder der Neuansatz von Äußerungen sind

übliche Phänomene: Sätze sind zum Teil unvollständig oder beginnen inmitten einer Äußerung neu, werden mittendrin korrigiert oder nachträglich durch Zusatzinformationen ergänzt. Häufig fehlen wesentliche Informationen wie z. B. Nomen in der Äußerung, die nur aus dem Kontext generierbar sind. Einzelne Phrasen werden verwendet, die oftmals nicht in einen vollständigen Satz eingebettet sind. Ebenfalls werden einzelne Wörter inmitten einer Äußerung wiederholt sowie Partikel wie „ähm“ oder Füllwörter wie „ruhig“ oder „halt“ eingeschoben (siehe Dialogausschnitt 6.4). Verschleifungen von Wörtern sind üblich, was die Erkennung der korrekten Äußerung erschwert. Besonders deutlich wird der Unterschied der Spontansprache zu geschriebenen Texten im Dialogausschnitt 6.5.

Person: ähm, rechts drehen und einen Schritt vor bis –
ähm ja nach unten greifen und den Feuchtigkeitsgrad fühlen
Roboter: FEUCHT
Person: ähm, unten im – im Blumentopf nachgucken
Roboter: IST TROCKEN
Person: ähm, geringe – geringe Mengen Wasser nachfüllen
Roboter: OK
Person: gut, ähm diese Blume ist damit abgehakt
Roboter: OK
Person: Drehung – hundertachtzig Grad – vorgehen bis zum Schreibtisch
Roboter: OK
Person: Position drehen kurz nach rechts – Pflanze steht jetzt
direkt vor dir – Feuchtigkeit prüfen bitte
Roboter: OK
Person: trockener?
Roboter: FEUCHTER
Person: wenige Mengen Wasser hinzufügen
Roboter: OK
Person: *(*leise zu sich selbst:) a das ist wahrscheinlich eher mhm ok –
mhm – ja dann haken wir auch diese Pflanze ab und gehen
zur nächsten – mhm
⋮
Person: Programm abgeschlossen

Abbildung 6.4.: Dialog-Beispiel: Programmähnliche Instruktionen

Der Aspekt der Situiertheit ist in den Dialogen ein wichtiges Merkmal. 13 von 14 Personen verwendeten Zeigegesten in den Dialogen, um auf Objekte zu referieren (diese sind in den Abbildungen mit $< \ldots >$ dargestellt). Dabei treten die Gesten in der Regel zusammen mit deiktischen Äußerungen wie „das da“ oder „dieses“ auf. Im Dialogausschnitt in Abbildung 6.6 ist die Kombination von Gesten mit deiktischen Konstrukten besonders deutlich zu erkennen. Diese auf die Szene referierenden Äußerungen können nur in Verbindung mit Wissen aus der Umgebung inter-

Person: folgende Blumen – die die Blume mit dem blauen Blumentopf
Roboter: AUF WELCHEM TISCH ?
Person: ah – auf dem Tisch vor der Tafel
Roboter: OK
Person: dann die äh, die zweite große Blume – mit dem weißen Topf
die dort steht – vor dem Fenster
⋮
Roboter: DIE HINTERE ?
Person: die neben – die vor vor – auf dem Tisch vor dem äh –
vor dem Tisch mit dem großen Blumentopf steht
Roboter: OK
⋮

Abbildung 6.5.: Dialog-Beispiel: Ausgeprägte Spontansprache

pretiert werden. Die Verarbeitung multimodaler Informationen ist daher eine wichtige Aufgabe für eine gelungene Interaktion zwischen Mensch und Roboter. Ebenfalls verweisen die Probanden auch auf Objekte im Raum mittels räumlicher oder attributiver Zusatzinformation wie in Abbildung 6.7. Kontextinformationen aus der Umgebung sind notwendig, um die Äußerungen verstehen zu können. Gesten und zusätzliche Attribute wurden oft in Kombination verwendet wie in Abbildung 6.8).

Person: gucke <diese Blume> – einmal pro Woche gießen
Roboter: OK
Person: < diese Blume >
Roboter: WELCHE?
Person: < diese >
Roboter: DIE IM WEISSEN TOPF?
Person: Ja
Roboter: OK
Person: Auch einmal in der Woche gießen
Roboter: OK
Person: < gucke diese Blume >
⋮

Abbildung 6.6.: Dialog-Beispiel: Kindliche Anweisung

Auffällig war auch, dass die verschiedenen Probanden ein breites Spektrum an Dialogstilen wiedergaben. Einige ähnelten der Interaktion mit einem Kind wie in Abbildung 6.6, andere waren stark angelehnt an einem Programmierstil wie in Abbildung 6.4. Beispielsweise verwendeten einzelne Personen unerwarteter Weise sehr formelle Ausdrücke wie z. B. „zehn Milliliter Wasser“ anstelle von „nur wenig Wasser“, wie in dem Dialogausschnitt in Abbildung 6.2 dargestellt.

Person: und hier vorne auf dem Tisch
Roboter: OK
Person: kleiner weißer Topf
Roboter: OK
Person: Mittwoch Mittag - zehn Milliliter
⋮

Abbildung 6.7.: Dialog-Beispiel: Kontextinformation

Person: < die hier vorne auf dem Schreibtisch >
Roboter: WELCHE?
Person: < diese hier vorne >
⋮

Abbildung 6.8.: Dialog-Beispiel: Referenz durch Zeigen

Ebenfalls gab es einige Probanden, die einen sehr freien Kommunikationsstil führten, zum Teil auch mit umgangssprachlichen Redewendungen wie die Dialoge in Abbildung 6.3 und 6.5 zeigen.

Fragen tauchten aufgrund der Art der Dialoge selten auf. Aber selbst hier zeigt sich, dass sie zum Teil nur aus dem Kontext oder durch zusätzliche Prosodieinformationen zu erkennen sind. Im Dialogausschnitt in Abbildung 6.4 äußert der Proband „trockener“. In diesem Fall ist das eine Frage an den Roboter, die nur durch die Intonation zu erkennen ist.

Im Gegensatz zu dem in [Pet99] beschriebenen Korpus verwenden die Probanden in diesem Szenario Höflichkeitsfloskeln dem Roboter gegenüber. Womöglich kann es daran liegen, dass der Roboter von einem echten Menschen simuliert wurde. In den realen Experimenten mit dem Robotersystem zeigt sich jedoch, dass auch hier Höflichkeitsformeln verwendet werden und auch soziale Interaktionen stattfinden („Hallo“, „Wie heißt du?“, „bitte“ usw.).

6.2. Das englische Korpus: *Hometour*

Ein weiteres Szenario für die Mensch-Roboter-Interaktion ist das *Hometour-Szenario*. Der Gedanke dieser Interaktion ist die Ausgangskonstellation eines Robotersystems, das neu in die Umgebung eingeführt wird. Nach der Ankunft im neuen Heim werden ihm die für ihn noch unbekannten Räume und Objekte vorgestellt, die für die weiteren Handlungen des Roboters relevant sind. In Abbildung 6.9 ist eine typische Situation dargestellt. Das *Hometour-Szenario* ist im Rahmen des europäischen Projektes COGNIRON entstanden, das sich mit Forschungsfragen zu kognitiven Fähigkeiten von Roboter-Gefährten befasst. Die verwendete Sprache ist in diesem Szenario Englisch.

Abbildung 6.9.: Eine typische Interaktionssituation im Hometour-Szenario.

Zunächst einmal wurden theoretische Überlegungen angestellt, um die potentiellen Dialoge zu beschreiben. Dabei wurde zusätzlich das bereits vorhandene Wissen aus den Experimenten im Rahmen des *Blumengieß-Szenarios* genutzt. Die Dialoge sollten zusätzlich zu dem bereits vorhandenen Korpus des *Gieß-Szenarios* den Wortschatz für den Roboter bilden. Dafür wurde das Lexikon angepasst, das zum einen Übersetzungen des deutschen Lexikons beinhaltete und zusätzlich die Ergänzungen aus dem *Hometour-Korpus* erhielt. Die semantischen Entitäten wurden neben den schon existierenden Entitäten aus dem *Gieß-Szenario* ebenfalls um die neu gewonnenen aus dem *Hometour-Korpus* erweitert. Erst aufgrund der theoretisch erstellten Dialoge und der daraus resultierenden Datensammlung konnte der Roboter seine sprachlichen Fähigkeiten bilden und mit den Interaktionspartnern kommunizieren.

Die Dialoge für das *Hometour-Korpus* wurden anhand verschiedener Kriterien erstellt. Zum einen bilden die Rahmenbedingungen aus dem *Gieß-Szenario* Richtlinien, nach denen sich auch die künstlich ausgedachten Dialoge richten sollten. Das sind u. a. soziale Interaktionsfähigkeiten, die Einbeziehung der Situation in den Dialogen und die Bereitstellung einer möglichst freien Kommunikation. Zum anderen sind die Eigenschaften und Fähigkeiten, die der Roboter besitzt und besitzen soll, ein wichtiger Aspekt. Folgende Fragestellungen wurden bei der Erstellung der potentiellen Dialoge berücksichtigt: Wie könnte eine Interaktion mit einem Roboter aussehen? Welche Fähigkeiten könnte der Roboter besitzen, die kommuniziert werden können? Auf welche sprachlichen Besonderheiten sollte der Roboter eingehen können? Welche Probleme könnten in der Interaktion auftreten und wie kann der Roboter möglichst sinnvoll damit umgehen? Welche Objekte oder Gegenstände können Thema der Interaktion sein und was sind wichtige Unter-

scheidungsmerkmale für den Roboter? Es ist z. B. für den mobilen Roboter wichtig, dass er bewegliche Objekte wie eine Tasse von festen Gegenständen wie einem Fenster oder einer Tür unterscheiden kann, da diese zur räumlichen Orientierung dienen können. Ebenso kann das Wissen um die Funktion oder Kategorie eines Objektes hilfreich für die weitergehende Interaktion sein. Ausschnitte aus Beispieldialogen sind in 6.10 und 6.11 abgebildet.

Person: What is your name, robot?
Robot: MY NAME IS BIRON.
Person: What can you do?
Robot: I CAN FOLLOW YOU AND YOU CAN SHOW ME SOMETHING.
Person: Please follow me.
Robot: OK. <FOLLOWING>
Person: (Look,) this is a cup, it is green.
Robot: OK, THE CUP IS GREEN.
Person: ???
Robot: I BEG YOUR PARDON?
Person: Follow me.
Robot: OK.

Abbildung 6.10.: Dialog-Ausschnitt aus dem theoretisch erstellten *Hometour-Szenario*

(walking to the bedroom)
Person: Stop, this is my bedroom.
Robot: IT LOOKS COMFORTABLE.
Person: Thank you. Look, this is my bed.
Robot: SORRY, I CAN'T SEE IT, MAYBE IT IS TOO DARK.
CAN YOU TURN ON THE LIGHT?
Person: Oh, sorry. (turns on the light) Can you see the bed now?
Robot: YES, THANK YOU.
Person: Look, that is a bedside table, it is white as the bed.
Robot: IT IS REALLY NICE.

Abbildung 6.11.: Dialog Ausschnitt aus den theoretisch erstellten *Hometour-Szenario*

Für das Sprachverstehen wurden die Dialoge nach verschiedenen Kriterien untersucht: Welche Dialogakte sind relevant für eine erfolgreiche Kommunikation? Welche sprachlichen Phänomene spiegeln sich in den Beispieldialogen wider? Ebenso wurden das Lexikon und die semantische Datenbank durch die Dialoge erweitert (siehe auch Kap. 8).

Als erstes wurden die Aufgaben festgelegt, die der Roboter können soll oder bereits kann. Das sind in erster Linie folgende Aufgaben: das Lernen seiner Umgebung und darin befindlicher Objekte, das Folgen von Personen, Botendienste sowie das Herumführen anderer Personen durch die Räumlichkeiten. Zusätzlich zum *Hometour-Korpus* wurden weitere Kontexte festgelegt, die ebenfalls Thema eines mobilen Robotersystems sein können. In diesen Kontexten könnten beispielsweise folgende Äußerungen vorkommen:

- Hello Robby, I would like to show you your recharge station for your batteries. Please follow me.
- Hi Rob, please clean this window.
- BIRON, get the garbage to the container.
- Clean the table surface. You will find a cloth in the kitchen.
- Turn down the heater/radiator. Regulate the heater regulator down to position three.
- Get me the post/newspaper.
- Please show our new collegue around this place.
- Please fetch me a cup of coffee, sugar and milk as usual.

Ein weiterer wichtiger Bereich sind die Objekte, die der Roboter erkennen und über die er kommunizieren können soll. Zuerst einmal sind das feste Bestandteile des Gebäudes, wie Fenster, Türen oder Raumkonzepte, desweiteren sind das feste Einrichtungsgegenstände wie Möbel oder Elektrogeräte. Auch frei bewegliche Gegenstände gehören zu den Objekten, über die der Roboter kommuniziert wie z. B Geschirr, Büroutensilien und Lebensmittel. Zusätzlich wurden die Eigenschaften erfasst, die diese Objekte besitzen können.

Besonders wichtig für die korrekte Erfassung der Dialoge und für die kommunikativen Fähigkeiten des Roboters sind die Dialogakte, die in dem Szenario vorkommen. Diese Dialogakte können das Dialogsystem unterstützen, schnell und angemessen zu reagieren. Im Gegensatz zu dem in [Fis95, BP99a] beschriebenen Korpus von Konstruktionsanweisungen weisen die Dialogakte in dieser Arbeit eine wesentlich allgemeinere Form auf. Sie sollen eine Vielzahl an Themen abbilden und beschränken sich nicht auf den Bereich der Konstruktionsanweisungen. Folgende Dialogakte wurden in dem *Hometour-Szenario* der Mensch-Roboter-Interaktion unterschieden:

- *Socialisation*: „Sorry“, „Thank you“, „Good-bye“, „Hello“, ...
- *Instruction*: „Follow me“, „I will show you *X*“, ...
- *Query*: „What can you do?“, „Where is *X*?“, ...
- *Confirmation*: „Yes“, „OK“, „Good“, ...

Phänomene	Kontext	Beispiel
Spontansprache	(immer möglich)	„the flower ehm, the red one“
Situiert: Geste und Sprache	(Objekt zeigen)	„the red thing < over there >“
Räumliche Relationen	(Objekt bestimmen)	„the cup next to the teapot“
Themenwechsel	(neues Diskurssegment)	„hello – follow me – look here“

Tabelle 6.1.: Sprachliche Phänomene

- *Negation*: „No“
- *Deletion*: „Forget it“
- *Correction*: „Not *X*, but *Y*“
- *Description*: „This is *X*“
- *Object*: Möbel, Geschirr, Lichtschalter, Bürogegenstände, ...
- *Fragment*: (alle Wörter, die keinem anderen Dialogakt zugeordnet werden können)

Aus dem Korpus konnten verschiedene sprachliche Phänomene detektiert werden, die in Tabelle 6.1 dargestellt werden. Ein weiteres Thema ist die Auflösung von Anaphern. In realen Szenarien ist nicht immer eindeutig, ob es sich um eine rein sprachliche Anapher handelt, oder ob auf ein Objekt in der Szene referiert wird. Anapherresolution ist demnach stark kontextabhängig. In den meisten Fällen kann eine anaphorische Äußerung nur durch zusätzliches Umgebungswissen aufgelöst werden. Dabei ist hilfreich, wenn der Hinweis auf benötigtes Szenenwissen aus sprachlichen Ressourcen kommt. Nur wenn die Äußerung nicht durch Umgebungsinformationen gelöst werden kann, handelt es sich vermutlich um eine rein sprachliche anaphorische Verwendung (vgl. Kap. 9.4).

6.3. Das englische Korpus: Experimentdaten

Die Experimente im Rahmen des *Hometour-Szenarios* entstammen der Idee, eine möglichst freie Interaktion ohne Einschränkungen zu ermöglichen. Es taucht ebenfalls die Frage auf, was sowohl der Dialog als auch das Gesamtsystem leisten müssen, damit eine sinnvolle Interaktion stattfinden kann und der Roboter von den Benutzern akzeptiert wird. Besonders wichtig für das Sprachverstehen war der Abgleich zwischen dem, was die Probanden tatsächlich äußerten, im Vergleich zu den nach theoretischen Überlegungen erstellten Dialogen. Die gesamten Experimente fanden mit dem Robotersystem BIRON im laufenden Betrieb statt. Das bedeutet, dass eine reale Kommunikationssituation gegeben war und alle Systemkomponenten aktiv waren. Die genaue Beschreibung der Experimente findet in Kapitel 10.3 statt.

Während der Experimente konnten die Probanden mit dem Roboter frei kommunizieren, es wurden keinerlei Einschränkungen gemacht, was sie sagen sollten. Um ihnen den Einstieg möglichst einfach zu machen, wurde ihnen gesagt, sie könnten den Roboter fragen, welche Fähigkeiten er besitzt, und diese dann ausprobieren. Während des Experimentes war das gesamte Robotersystem mit all seinen Funktionalitäten aktiv, das heißt, der Roboter erhielt neben gesprochener Sprache u. a. auch visuelle Informationen über seine Interaktionspartner und seine Umgebung. Die Kamera wurde durch die Aufmerksamkeitssteuerung in Richtung der Versuchsperson gelenkt. Die gesamte Interaktion sowie auch die Experimentbeschreibung liefen in englischer Sprache ab, um schon im Vorfeld ein Umdenken auf diese Sprache zu erleichtern.

Problematisch bei dem Experiment war die sehr freie Gestaltungsmöglichkeit der Interaktion. Die Möglichkeiten, die die Benutzer äußerten, überstiegen bei weitem die Fähigkeiten des Spracherkenners, der einen stark eingeschränkten Sprachraum besitzt. Da die Probanden mit dem Roboter auf Englisch kommunizieren sollten, die meisten jedoch deutsch als Muttersprache sprachen, und zusätzlich Störgeräusche die Interaktion beeinflussten, stand die Spracherkennung vor besonderen Herausforderungen und war besonders fehlerträchtig. Die Partizipanten wurden häufig falsch verstanden, so dass der Roboter auf viele Äußerungen nicht adäquat reagieren konnte. Ebenfalls fragten die Probanden Dinge oder erwarteten Handlungen, die das Robotersystem überforderten und die es somit nicht beantworten oder ausführen konnte.

Doch gerade durch die freie Gestaltung konnten sehr interessante Beobachtungen gemacht werden. Wie auch in der ersten Studie äußerten die Versuchspersonen eher kurze Sätze. Sie machten häufig Pausen mitten in der Äußerung, vor allem, wenn Gesten involviert waren (siehe Abb. 6.12 und 6.13). Die Äußerungen zeigten spontansprachliche Phänomene wie Hesitationen, Auslassungen oder Wortwiederholungen. Ebenfalls entsprachen nicht alle Äußerungen den Grammatikregeln der geschriebenen Sprache (siehe Abb. 6.14). Umgebungswissen wurde häufig in die sprachlichen Aussagen integriert und Zeigegesten verwendet (neben anderen Modalitäten wie z. B. Mimik).

Person1: Can you see the cup?
Look – < this – is a cube – this > – BIRON.
Follow me. Stop.
Robot < look > – do you see?
BIRON < look >.
< This is a cow > – funny.
Do you like it?
What –
< Look – this – is a puncher >.
That‘s fine.

Abbildung 6.12.: Dialog Ausschnitt aus dem *Hometour*-Experiment

Viele Personen änderten ihren Dialogstil während der Interaktion, je nachdem wie viel der Roboter verstand. Zuerst redeten sie recht normal wie auch mit einem menschlichen Gegenüber. Wurden Äußerungen mehrfach nicht verstanden und mussten die Probanden sie mehrmals wiederholen, über-artikulierten sie einzelne Worte und wurden auch lauter. Einige Personen wechselten in einen Baby-Sprachstil (engl. *baby-talk*), beispielsweise in Dialog 6.12: „BIRON - look - do you see? BIRON look. This - is a cow - funny. Do you like it?“

Person2: Hello BIRON.
Please don‘t tell me it‘s my fault.
I said Hi, BIRON.
OK. Glad to hear that.
Hi, BIRON, what – can you do?
That‘s great.
< This – is a book > – is a book.
< Look here >.
And look < here > – a cup – a cup.
...
< Look is a cup – cup > .
So good – < this – is a cup >.
We are getting somewhat.

Abbildung 6.13.: Dialog Ausschnitt aus dem *Hometour*-Experiment

Die Benutzer äußerten häufig Meta-Kommentare, die die momentane Situation beschrieben und kommentierten wie „that‘s fine“ oder „that’s great“ (siehe Dialogausschnitt 6.12 und 6.13). Sie kommentierten ihre Bewertungen über das Roboterverhalten und die Situation. Diese Meta-Kommentare sind teilweise aufgrund der Antworten des Roboters selbst entstanden. Der Roboter kommentiert, wenn er etwas nicht verstanden hat, beispielsweise damit, dass es ihm leid tue, dass er das nicht verstanden habe. Bei erkannten Objekten sagt er auch, dass er sie mag[1]. Meta-Kommentare kamen auch in Situationen vor, wenn die Probanden den Roboter selbst falsch verstanden haben wie z. B. der Kommentar in Dialogausschnitt 6.13: „Please don‘t tell me it‘s my fault.“, wo der Roboter sich für seine schlechte Spracherkennung entschuldigt hat, der Proband sich selbst aber verantwortlich gemacht fühlt. Teilweise wurden Meta-Kommentare der Probanden auch durch die schlechte Spracherkennung provoziert, z. B. Kommentare wie „We are getting somewhat“ . Überdies machten sie auch Bemerkungen auf Deutsch, sie wandten sich an den Experimentator oder beschwerten sich über die schlechte oder missglückte Kommunikation.[2] All diese Kommentare wurden im Vorfeld nicht erwartet und konnten vom System daher auch nicht interpretiet werden.

[1] Kommentare des Roboters:„I really like it“, „I know it is sometimes difficult with me, but don’t feel discouraged!“, „It is my fault, that ...“.

[2] Beispiele für deutsche Bemerkungen: „also das geht“, „Blöde Kiste!“oder „Ach Baby, du bist ja heute mal gut in Form“

Person3: Let me show you another one –
show you a cube. *(nimmt Würfel hoch)*
Follow me. Stop.
< This – is a book >.
< This is blue cup >.
Bye.

Abbildung 6.14.: Dialog Ausschnitt aus dem *Hometour*-Experiment

Person4: Can you see me?
Can you follow my finger? *(bewegt Finger)*
Oh that's not so nice.
Can you hear music?
Can you walk in this room?
Sorry, can you repeat your answer?
How fast can you move?
Please stop here.
Stop moving – fine.
What is your name?
How old are you?
Sorry, can you repeat your answer?
Have you got any hobbies?
Ok, I understand.

Abbildung 6.15.: Dialog Ausschnitt aus dem *Hometour*-Experiment

Unerwarteterweise richteten einige Probanden viele Fragen an den Roboter, die vielleicht einem Menschen gestellt werden, aber bei einem technischen Gerät so nicht zu erwarten waren, z. B. die Frage nach den Hobbys oder dem Alter des Roboters. Bei den Fragen ist ebenfalls zu bemerken, dass nicht immer zu erkennen ist, ob es sich um eine Frage oder eine Anweisung handelt. Probanden sind mitunter höflich und verwenden indirekte Sprechakte anstelle von direkter Aufforderung wie in Dialogausschnitt 6.15 zu sehen ist. Nur aus dem Kontext heraus kann erkannt werden, ob die Frage „can you walk in this room“ oder „can you repeat your answer“ eine Frage oder eine Aufforderung ist.

Anhand dieser Experimente wurde der Sprachumfang für das Sprachverstehen des Roboters erweitert und adaptiert. Zu großen Teilen wurden jedoch auch die theoretischen Überlegungen des *Hometour-Korpus* mit einbezogen. Die Dialogakte scheinen stimmig zu sein mit den Aussagen der Probanden. Ebenfalls nutzten die Probanden Umgebungsinformationen und verwendeten Zeigegesten für Objektreferenzen. Insgesamt waren die Äußerungen eher kurz, enthielten jedoch Merkmale der Spontansprache.

6.4. Zusammenfassung

Es ist schwierig, allgemeine Aussagen zu machen über das generelle Kommunikationsverhalten von Menschen gegenüber Robotersystemen. Es gibt kaum Studien in diesem Bereich. Einerseits kann Mensch-Mensch-Kommunikation nicht direkt auf Mensch-Roboter-Kommunikation übertragen werden, da sich Menschen schnell an die Fähigkeiten und den Sprachstil des Gegenübers anpassen. Ihre Vorerfahrungen und inneren Erwartungen beeinflussen ihre Kommunikation. Daher weisen die Benutzer auch sehr unterschiedliche Kommunikationsstile auf, die von einer sehr freien Unterhaltung bis hin zum Programmierstil reicht. Andererseits sind die Themen und Inhalte weit gestreut und nicht alle Äußerungen sind im Rahmen dessen, was im Vorfeld erwartet werden kann. Diese können dann auch nicht modelliert werden und Missverständnisse zwischen Mensch und Roboter sind daher kaum zu vermeiden.

Dennoch sind im Kommunikationsverhalten der Probanden generelle Tendenzen zu erkennen, die vom System berücksichtigt werden sollten. Die Äußerungen der Benutzer sind tendenziell kurz, Nebensätze werden selten geäußert, vereinzelt bestehen die Äußerungen aus Satzfragmenten (vgl. [BP99a] S.41). Dagegen enthalten sie spontansprachliche Artefakte, wie sie zum Teil auch in [Kro00, Pet99] beschrieben wurden. Der traditionelle Satzbegriff kann in dieser Domäne nicht die Form der Äußerungen beschreiben.

Neben dem Aspekt der Spontansprache ist die Situiertheit ein weiterer zentraler Aspekt des Korpus. In den Dialogen wird regelmäßig auf die Umgebung referenziert, was sich zum Teil in der Sprache bemerkbar macht, z. B. durch deiktische Äußerungen. Diese Äußerungen können nur vollständig interpretiert werden, wenn Kontextinformationen aus der Umgebung berücksichtigt werden, u. a. durch die Einbindung einer Gesten- und Objekterkennung (siehe auch Abschnitt 2.3). Andererseits können sprachliche Hinweise auf visuelle Informationen in der Szene auch die Verbindungen zu anderen Modalitäten stützen.

Insgesamt trägt die Korpusanalyse dazu bei, den Sprachumfang des Robotersystems festzulegen und überdies, die Anforderungen und die Designkriterien an das sprachverstehende System zu erkennen (siehe Kap. 7). Zusammenfassend beschreiben folgende Merkmale die Kernpunkte der Dialoge zwischen Mensch und *Robot Companion*:

- Die Äußerungen weisen deutliche Merkmale von *Spontansprache* auf. Sowohl im Deutschen als auch im Englischen ist die Wortfolge recht frei, die Äußerungen der Probanden sind tendenziell kurz, bilden jedoch ein breites Spektrum an komplexen Phänomenen wie Hesitation, Neuanfang, Wiederholungen usw. ab.
- Die Eingebundenheit der beteiligten Interakteure in der Szene wird durch die *situierte Sprache* deutlich. Probanden referieren auf Objekte durch Gesten oder indirekt durch sprachliche Referenzen. Syntaktisch zeigt sich die Situiertheit durch eine vermehrt deiktische und anaphorische Sprache und durch Auslassungen. Die Interaktion in einer dynamischen Umgebung ist sehr anspruchsvoll und aufwendig und bedarf ständiger Adaptionsleistung.

- Probanden verwenden eine Vielzahl an inhaltlichen Themen, die eine große Bandbreite an Äußerungsmöglichkeiten umfassen. Der *Sprachwortschatz* des Roboters muss daher entsprechend groß sein.
- Die Probanden besitzen scheinbar keine einheitlichen Annahmen über die Fähigkeiten und Eigenschaften des Robotersystems. Das zeigt sich daran, dass die Probanden eine große Bandbreite an unterschiedlichen *Dialogstilen* verwenden.
- In den Dialogen tauchen immer unerwartete Äußerungen auf, bei denen das System nicht in der Lage ist, sie zu verarbeiten und an seine *Grenzen* stößt. Hier können eine flexible Erweiterbarkeit der Sprachverarbeitung sowie die Darstellung der internen Zustände des Robotersystems von Vorteil sein.
- Die Probanden sehen den Roboter in den meisten Fällen als sozialen Interaktionspartner an und verwenden daher *Höflichkeitsformen*.
- Die *Spracherkennung* kann nie hundertprozentig fehlerfrei sein, die Benutzer reagieren ihrerseits wiederum auf eine missglückte Kommunikation z. B. mit Rückfragen oder Kommentaren.

7. Anforderungen und Designkriterien für das Sprachverstehen

Die Anforderungen, die an ein Robotersystem mit sozialen Fähigkeiten gestellt werden, bilden ein breites Spektrum ab. Dabei muss sich auch die Sprachverarbeitung einigen Herausforderungen stellen, die speziell in diesem Kontext zu beachten sind. Die zentralen Aspekte für das Sprachverstehen werden im Folgenden kurz erläutert (vgl. [Hüw04]).

7.1. Verarbeitung situierter und spontaner Sprache

In Kapitel 6 wurde auf die Besonderheiten von situierten und spontansprachlichen Äußerungen ausführlich eingegangen. Diese Eigenheiten stellen besondere Anforderungen an das Sprachverstehen. In [Pot99] wird ebenfalls betont, dass die Verarbeitung von Spontansprache ein wichtiger Aspekt für benutzerzentrierte Dialogsysteme ist.

Ein besonderes Merkmal ist, dass in den Äußerungen viele Freiheitsgrade existieren. Die Anwender versprechen sich, wiederholen Wörter, brechen Äußerungen ab oder verbessern sich selbst. Die Möglichkeiten sind dabei vielfältig. Möchte man spontansprachliche Dialoge verarbeiten, muss man u. a. mit den Phänomenen der Disfluenz umgehen können [McK98]. Häufig vermischen sich die rein sprachlichen Informationen mit den Informationen aus der Szene, was ebenfalls in vielen Äußerungen sichtbar wird. Situierte Sprache erfordert besondere Anforderungen an die Verarbeitungskomponenten. Zudem ist es schwierig, vorherzusagen, was die Interaktionspartner dem Roboter mitteilen. Aufgrund unserer Erfahrungen während der Experimente gehen wir daher eher von der Verarbeitung eines großen Sprachumfangs aus, um möglichst viele Inhalte abzudecken. Dennoch muss damit gerechnet werden, dass die Partner unbekannte Wörter oder Phrasen verwenden.

Daher ist es nicht für alle Äußerungen möglich, sie vorher vollständig in eine syntaktische Struktur abzubilden. Zum einen lassen sich viele Äußerungen nicht vorhersagen und aufgrund dessen nicht abbilden. Zum anderen erscheint der Aufwand, spontansprachliche Äußerungen mit ihren Besonderheiten grammatikalisch abzubilden, recht hoch. Es ergeben sich mitunter Komplexitätsprobleme, weil die anfallenden Datenmengen groß und die kombinatorischen Möglichkeiten bei der Wort-Satzbildung enorm sind (siehe Abschnitt 7.5). Kronenberg [Kro00] beschreibt einen Parser für syntaktisches Parsen von spontansprachlichen Anweisungen, dabei wird deutlich, welch hohe Anforderungen an eine spontansprachliche syntaktische Analyse besteht. Leider gibt diese Arbeit keinerlei Hinweise auf den Datenumfang und den Umfang der grammatikalischen Regeln.

7.2. Robustheit

Bei der Interaktion in einer realen Umgebung wird das System auch immer fehlerbehaftete Daten erhalten und dementsprechend robust muss der Verarbeitungsmechanismus sein. Der Spracherkenner ist sowohl mit Störgeräuschen aus der Umgebung konfrontiert, z. B. Sprachsignalen von anderen Personen oder Türknallen, als auch mit den Geräuschen, die der Roboter aufgrund seiner eigenen Mechanik produziert. Im Gegensatz zu geschriebener Sprache besteht gesprochene Sprache aus einem akustischen Signal, das erst vom Spracherkenner in eine Art Wortkette umgewandelt werden muss. Fehl-Erkennungen sind daher nie auszuschließen. Umso wichtiger ist es, dass der Analysemechanismus dieses in seine Verarbeitung mit einbezieht. In [Kro00] ist eine Liste von spontansprachlichen Äußerungen und im Vergleich dazu im vom Spracherkenner erkannten Wortketten abgebildet, die sehr anschaulich demonstriert, wie stark diese voneinander abweichen können und wie wichtig daher eine robuste Verarbeitung ist. Diese Beispiele zeigen auch, wie wichtig es ist, sprachliche und visuelle Informationen zu verbinden, um das System gegenüber Fehlern sowohl in der Erkennung der Sprache als auch in der Erkennung von Bildinformationen abzusichern und um die Robustheit des Gesamtsystems zu erhöhen.

In vielen Fällen kann syntaktisches Parsen problematisch oder sehr aufwendig sein, wenn z. B. statt eines Verbs ein Nomen erkannt wurde. Auch häufen sich gerade zu Beginn und gegen Ende einer Äußerung die Fehler der Erkennung, weil nicht immer richtig erkannt wurde, wann eine Person anfängt zu sprechen und wann sie aufhört. Geräusche aus der Umgebung werden dann als Bestandteil der Äußerung fehlinterpretiert. Zudem kann es immer wieder vorkommen, dass Personen dem System unbekannte Wörter verwenden und diese dann missverstanden und auf andere Wörter abgebildet werden. Für die Analyse der Sprachinformationen liegt hier ein Unsicherheitsfaktor vor, der im Prozess berücksichtigt werden sollte.

Da Spracherkennung und Analyse meist verschiedene, unabhängig voneinander entwickelte Systeme sind, ist nicht unbedingt davon auszugehen, dass die Lexika der beiden Systemkomponenten identisch sind. Daher kann es vorkommen, dass ein Wort zwar richtig erkannt werden konnte, der Sprachanalyse jedoch unbekannt ist. Diese unbekannten Wörter sollten nicht zum vollständigen Abbruch einer Analyse führen.

Der Verarbeitungsmechanismus sollte robust gegenüber Erkennungsfehlern sein. Da diese Fehler in situierten Kontexten vermehrt auftreten, ist es wichtig, dass die Sprachverarbeitung darauf eingeht. Sie sollte so viele Informationen wie möglich aus der erkannten Äußerung gewinnen können, um die Interaktion insgesamt so natürlich wie möglich zu gestalten. Erst dann kann auch eine fruchtbare Kommunikationsbeziehung zwischen Mensch und Roboter entstehen.

Mehr noch ist es äußerst hilfreich, wenn die Korrektheit der erkannten Äußerung ebenfalls berechnet werden kann. Diese liefert hilfreiche Hinweise für den Dialog. Ist die Äußerung korrekt verstanden worden, kann entsprechend der Intention des Interakteurs reagiert werden. Wurde die Äußerung nur in Teilen verstanden, kann das ebenfalls eine Stütze für den Dialogmanager sein. Er kann gezielt Rückfragen stellen, sich auf die richtig verstandenen Teile beziehen oder in der Umgebung nach ergänzenden Informationen suchen und so einen Abgleich zwischen Szene und

Sprache ausführen. Wurde die Äußerung insgesamt nicht richtig verstanden, kann er dann noch einmal nachfragen oder auch darum bitten, andere Wörter zu verwenden, um die Erkennungsrate zu verbessern.

Die Frage ist nun, wie das Sprachverstehen zwischen falsch erkannten und richtig erkannten Äußerungen unterscheiden kann. Gerade in der spontansprachlichen Situation sind Äußerungen nicht unbedingt wohlgeformt und daher kann die syntaktische Korrektheit keine Informationen darüber liefern, was die Interaktionspartner tatsächlich gesagt haben könnten. Im Gegensatz dazu, kann der Sinnzusammenhang, also die semantische Kohärenz, eine bessere Entscheidungshilfe über die Güte der Ergebnisse aus der Spracherkennung liefern.

7.3. Bereitstellung relevanter Informationen

Das Sprachanalysesystem muss vor allem semantische Interpretationen von Äußerungen liefern. Da das System jedoch in ein Robotersystem eingebunden ist, das mit Daten aus der realen Umgebung konfrontiert wird und neben Sprache auch visuelle Informationen verarbeiten muss, bestehen besondere Anforderungen an die Informationen, die von der Sprachverarbeitung generiert und weitergereicht werden müssen.

Semantische Informationen

Zuerst ist die Bereitstellung semantischer Informationen die Kernaufgabe der Sprachverarbeitung. Der Dialogmanager benötigt diese Informationen, um mit seinem Interaktionspartner kommunizieren zu können. Er muss wissen, was der Anwender gesagt hat, welche Aufgabe der Roboter ausführen soll oder welche Antwort er geben soll. Dabei soll eine möglichst kohärente semantische Interpretation generiert werden.

Informationen über den Dialogakt

Zusätzlich ist es hilfreich, Kontextinformationen über den Dialogakt zu erkennen und an den Dialogmanager zu transferieren. Dieser kann dann im Zweifelsfall die richtige Handlung ansteuern, er weiß, ob er eine Anweisung ausführen oder abbrechen soll, eine Frage beantworten kann oder seine sozialen Kompetenzen gefordert sind.

Einfacher und schneller Zugriff

Die Ausgabe der Sprachverarbeitung soll möglichst einfach zu verarbeiten sein. Sowohl der Dialogmanager als auch andere beteiligte Systemkomponenten wie die Themendetektion müssen einfachen und schnellen Zugriff auf die Daten haben, ohne Konvertierungen vornehmen zu müssen. Die Beschreibungssprache XML liefert hierfür eine gute Basis. Zudem verwendet die Roboterplattform BIRON das XML basierte Kommunikationsverfahren XCF (siehe Abschnitt 5.3) und stellt somit Funktionen für den einfachen Austausch der Daten bereit.

Multimodalität

Oftmals besitzen sprachliche Äußerungen einen Bezug zur realen Umgebung. Für eine gelungene Interaktion muss dieses Wissen aus der Sprache mit den Informationen aus der Umgebung zusammengeführt werden können. Sprache bietet direkte oder indirekte Hinweise auf visuelle Informationen die vom System für die Zusammenführung genutzt werden können. Mehr noch können Hinweise aus der Sprache die visuellen Komponenten des Robotersystems triggern und so wesentlich zur Einsparung von Rechenzeit beitragen. Hat eine Person „diese hier“ gesagt, kann daraus geschlossen werden, dass die Person eine entsprechende Geste auf ein Objekt gemacht hat. Das Robotersystem kann daraufhin die zeitlich dazu passende Bildsequenz aus seinem Speicher im Hinblick auf eine geeignete Geste und einem darauf verwiesenem Objekt analysieren. Auch können sowohl sprachliche als auch visuelle Informationen „verrauscht“ sein. Die Zusammenführung beider Modalitäten macht das Robotersystem robuster gegenüber unsicheren Daten.

Mit Hilfe eines so genannten Zeit-Stempels jeweils für den Zeitpunkt der Entstehung der Äußerung als auch für gewonnene Informationen aus der Szene, zum Beispiel aufgrund einer erkannten Zeigegeste auf ein Objekt, kann eine Verknüpfung dieser verschiedenen Modalitäten stattfinden. Dieser Zeit-Stempel wird von der Spracherkennung generiert. Sowohl die Hinweise auf Szeneinformationen als auch der Zeit-Stempel sollten an den Dialogmanager weitergereicht werden.

Bewertung der Interpretationsgüte

Nicht nur visuelle Bilddaten bilden im System einen Unsicherheitsfaktor, sondern auch die sprachlichen Eingaben. Unter Umständen bestehen Überlagerungen mit anderen Sprachsignalen oder Geräuschen, die in den Erkennungsprozess einfließen. Der Spracherkenner kann daher nie vollständig verhindern, dass Äußerungen aufgrund von Störgeräuschen falsch erkannt werden.

Um die Kommunikation mit dem Interaktionspartner zu erleichtern ist es von Vorteil, diese Unsicherheit mit zu bewerten. Dann kann der Dialogmanager bei dem Interaktionspartner nachfragen, wenn er meint, die Äußerung nur teilweise verstanden zu haben. Er kann aber auch gezielt die Informationen nutzen, die er verstanden hat und so eine natürlichere Kommunikation generieren. Eine Bewertung anhand syntaktischer Korrektheit abzugeben erscheint aufgrund der spontansprachlichen Erscheinung der Äußerungen nicht sinnvoll. Jedoch kann die semantische Kohärenz der interpretierten Äußerung dafür ein mögliches Bewertungskriterium liefern. Nach welchen Kriterien die semantische Kohärenz bestimmt wird, ist in Kap. 9.2.3 beschrieben.

7.4. Erweiterbarkeit und Adaptierbarkeit

Die Möglichkeit, die Wissensdatenbank der Sprachverarbeitung erweitern oder ändern zu können, ist gerade in der Entwicklungsphase ein zentrales Anliegen. Die Entwicklung eines erfolgreichen Benutzerinterfaces ist typischerweise ein iterativer Prozess, mit Designzyklen, Be-

nutzerstudien gefolgt von Designänderungen und Verbesserungen basierend auf den Evaluationen [Kam97]. Aus den Experimenten in realen Kontexten wurde deutlich, dass die Probanden oft unvorhergesehene Äußerungen und vor allem nicht berücksichtigte Inhalte verwendeten sowie Wörter nannten, die dem System unbekannt waren. In der Mensch-Roboter-Interaktion wird man immer wieder Äußerungen begegnen, die man so nicht bedacht hat. Denn erst durch die Experimente in realen Situationen in der Interaktion mit dem Roboter treten bestimmte Äußerungen auf, die im Vorfeld nicht berücksichtigt werden konnten, da nur wenige Vorerfahrungen und Annahmen über die Interaktion zwischen Mensch und Roboter-Gefährte existieren (siehe auch Kapitel 2). Um den Roboter möglichst intuitiv zu gestalten, müssen diese neuen Informationen in das System ohne großen Aufwand hinzugefügt werden können.

Auch nach Abschluss der Experimentalphase ist es wichtig, den Sprachumfang an die genaue Aufgabenstellung des Robotersystems anpassen zu können. Das Feedback der Benutzer kann entscheidende Hinweise für die Systemverbesserung geben [Pot99]. Die Möglichkeiten der Interaktion mit einem Robotersystem sind vielfältig. Weitere Aufgaben für den Roboter können hinzukommen und neue Anforderungen an das Robotersystem entstehen. Die Umgebung, in der sich der Roboter befindet, kann variieren. Das Sprachsystem sollte daher auf andere Kontexte adaptierbar sein und eine Möglichkeit bieten, Datensätze einfach austauschen zu können.

Aufgrund dessen ist es sinnvoll, eine Trennung zwischen Verarbeitungsmechanismus und Wissensdatenbanken vorzunehmen und dadurch das System möglichst flexibel und handhabbar zu gestalten. Auch sollten die Wissensdatenbanken relativ einfach zu erweitern sein.

7.5. Verarbeitungszeit und Effizienz

In der Verarbeitung von Sprache geht es darum, mit möglichst wenig Aufwand möglichst alle in der Sprache oder in dem Kontext auftretenden Äußerungen verarbeiten zu können. Wenn das System Minuten benötigt, um eine Antwort zu generieren, wird die Kommunikation sicherlich fehlschlagen [Ten03].

Während reguläre Sprachen in linearer Zeit geparst werden können, ist für kontextfreie Sprachen kubischer Aufwand erforderlich. Kontextsensitive Sprachen hingegen besitzen schon exponentiellen Aufwand (siehe [Bar87]). Offen ist bisher, ob natürliche Sprache mit kontextfreien Grammatiken beschreibbar ist: Zumindest das Zürichdeutsch ist nicht mehr kontextfrei. Den Beweis dafür liefert Shieber in [Shi85]. Im Gegensatz zu geschriebener Sprache ist die Grammatikalität von spontaner gesprochener Sprache deutlich geringer. Sie zeichnet sich durch Wiederholungen, Abbrüche und Korrekturen aus und lässt sich somit nicht mehr durch kontextfreie Grammatiken beschreiben (siehe [Kom95]). Daher besteht insbesondere in der syntaktischen Analyse die Problematik der Verarbeitungszeit. Um dieser zu entgehen, liefert der Ansatz semantische Verarbeitung eine handhabbare Lösung. Nicht die syntaktische Komplexität bestimmt die Verarbeitungszeit, sondern die Verarbeitung wird über den semantischen Zusammenhang gesteuert. Zusätzlich werden Annahmen über die Sprachstruktur gemacht (einfache Heuristiken), um die Komplexität der Verarbeitung der Äußerungen in der Domäne zu reduzieren.

7.6. Zusammenfassung und Fazit

Neben der Repräsentation von Wissen im Kontext der Domäne, ist die Verarbeitung dieser Daten für die Interaktionsfähigkeit des Roboters unumgänglich. Dabei existieren neben der Bereitstellung semantischer Informationen weitere Anforderungen, um einem kooperativen Dialog zwischen Mensch und Roboter adäquat gerecht werden zu können. Zum einen ist das eine schnelle Analyse, die sowohl in der Verarbeitungszeit den Anforderungen eines Echtzeitsystems gerecht werden kann, als auch wenig Speicherkapazität benötigt. Zum anderen muss dass System auf mögliche Fehler bei der Spracherkennung angemessen reagieren.

Zusammenfassend können folgende Anforderungen an die sprachverarbeitende Komponente für *Robot-Companions* angegeben werden:

- Verarbeitung von situierter und spontan gesprochene Sprache
- Verarbeitung und Bewertung von möglicherweise fehlerhaften Ergebnissen aus der Spracherkennung
- Schnelle Verarbeitung in „Echtzeit“
- Sprachumfang ist leicht erweiterbar und anpassbar
- Bereitstellung semantischer Informationen und Diskurskontext für den Dialog
- Bereitstellung von Hinweisen aus der Szene sowie Möglichkeit zur Verknüpfung der Daten mit Szenewissen bieten

Aufgrund der vielfältigen Anforderungen und Besonderheiten wird in dieser Arbeit ein Ansatz der rein semantischen Verarbeitung gewählt. Neben der Bereitstellung aller relevanten Informationen an den Dialogmanager und den anderen beteiligten Sprachkomponenten des Systems ist ein Ziel die Trennung von Verarbeitungsmechanismus und Wissensdatenbanken. Desweiteren kann mittels semantischer Kohärenz die Güte der Spracherkennung bewertet werden. Letztendlich werden in jedem Fall semantische Informationen benötigt und der Wegfall der syntaktischen Verarbeitung mit Fokus auf semantischer Verarbeitung stellt eine Alternative dar, die den Echtzeitanforderungen des Systems gerecht werden kann.

Im Folgenden werden zunächst die Daten beschrieben, die das Welt- und Diskurswissen des Roboters abbilden und anschließend der Mechanismus beschrieben, der diese Daten nutzt, um die Äußerungen der Interaktionspartner zu verarbeiten.

8. Wissensrepräsentation mit situierten semantischen Einheiten

Für das Verstehen von Sprache im Rahmen von Mensch-Maschine-Kommunikation ist das Wissen über die Domäne eine grundlegende Voraussetzung. Ohne Kenntnisse von semantischen Zusammenhängen in Form von internen Repräsentationen über Diskurs- und Weltwissen ist robustes automatisches Verstehen gesprochener Sprache auf einem mobilen Robotersystem nicht realisierbar. In diesem Kapitel wird das Konzept der Repräsentationen für mobile Robotersysteme dargestellt [Hüw06a, Hüw06b].

Die Herausforderungen, denen man sich bei der Planung und Implementierung eines Spontansprache verarbeitenden Systems für *Robot Companions* annehmen muss, wurden bereits in Kapitel 7 detailliert erörtert. Die in dem Kapitel aufgestellten Designkriterien betreffen nicht nur den eigentlichen Mechanismus, der die Äußerungen interpretiert und das Ergebnis an den Dialogmanager weiterleitet, sondern auch die Datenstrukturen, mit denen Diskurs- und Domänenwissen zur Verfügung gestellt werden. Diese Datenstrukturen müssen auf der einen Seite einfach und klar strukturiert sein, um die Verarbeitungsprozesse möglichst effizient realisieren zu können und um der wichtigen Anforderung der unkomplizierten Erweiterung des Sprachumfangs nachzukommen. Andererseits sollen die zentralen Aspekte der situierten Spontansprache abgebildet werden können. Diese beschränken sich nicht auf Sprache allein, sondern umfassen auch visuelle Informationen wie Gesten und Verbindungen von Sprache mit der aktuellen Szene. Darüber hinaus wurde beim Design der Datenstrukturen berücksichtigt, dass die vom System erkannten Äußerungen nicht nur syntaktisch, sondern auch semantisch unvollständig oder unverständlich sein können, nicht zuletzt auch durch mögliche Fehl-Erkennungen der Spracherkennung. Diese möglichen Fehl-Erkennungen müssen an den Dialogmanager weitergereicht werden, um entsprechend reagieren zu können. Daher sollen die Strukturen auch eine Infrastruktur bereitstellen, um die Ergebnisse der Spracherkennung bewerten zu können. Zusammenfassend lassen sich folgende Anforderungen für die Wissensrepräsentation im Kontext der Mensch-Roboter-Interaktion anführen:

- Die Wissensrepräsentationen müssen für die automatische Sprachverarbeitung geeignete Strukturen bilden und einen klaren Aufbau besitzen.
- Sie müssen alle relevanten Informationen im Bereich der situierten Spontansprache (semantische Relationen, Hinweise auf visuelle Informationen, Hinweise auf Dialogakt, usw.) darstellen können.

- Sie sollen einfach erweiterbar und adaptierbar sein, ohne Voraussetzung von Spezialwissen.
- Sie sollen eine Infrastruktur für die Abschätzung der Güte der Spracherkennungsleistung liefern.

In dem hier vorgestellten System wurde ein Lexikon integriert sowie situierte semantische Konzepte, die das benötigte Wissen für den Roboter bereitstellen. Sowohl Lexikon als auch die semantischen Konzepte, so genannte *situierte semantische Einheiten* („situated semantic units“ *SSUs*), sind speziell für Dialoge zwischen Mensch und mobilem *Robot Companion* konzipiert. Dabei wurde im Design eine klare Trennung vom lexikalischen Wissen und semantischem Wissen vorgenommen, um das Konzept der Erweiterbarkeit zu unterstützen. Damit können die unterschiedlichen Informationen unabhängig voneinander gespeichert und Redundanzen vermieden werden.

8.1. Lexikon

Das Lexikon enthält alle Wörter, die der Roboter benötigt, um die verbal an ihn gerichteten Aufgabenstellungen verstehen und interpretieren zu können. Dabei wurde aus Effizienzgründen die Variante der Vollform gewählt, um einen zusätzlichen Tagger einzusparen. Jeder Lexikoneintrag enthält einen Vollform-Namen, der der entsprechenden Ausgabe des Spracherkenners zugeordnet werden kann. Zusätzlich besteht jeder Lexikoneintrag aus syntaktischen Informationen, bei der zunächst einmal die jeweilige Wortart gespeichert wird, je nach Wortart und Bedarf ergänzt um weitere syntaktische Informationen. Für Nomen können das Numerus, Genus oder Kasus und für Verben Person, Tempus und Modus sein. Können einem Syn-Eintrag mehrere unterschiedliche Attribute zugeordnet sein, werden diese mit einem Balkenstrich getrennt aufgeführt. Da diese Informationen aber nur in den wenigsten Fällen für den hier eingesetzten Verstehensprozess von Nutzen sind, z. B. für Anaphernauflösung, wurde der vollständigen syntaktischen Zuordnung weniger Aufmerksamkeit geschenkt. Theoretisch ermöglichen die syntaktischen Informationen auch die Erweiterung des Parsingprozesses um syntaktische Strategien (vgl. Kap. 11).

Wesentlich wichtiger für das System ist die Zuordnung zu einem semantischen Konzept: Jeder Eintrag erhält dafür einen Verweis auf genau eine situierte semantische Einheit (SSU). Dies ist wichtig, um möglichst alle Wörter in die Verarbeitung integrieren zu können. Hätten nicht alle Lexikoneinträge eine semantische Konzeptrelation, würde der Verarbeitungsmechanismus einzelne Wörter einer Äußerung als semantisch unbekannt klassifizieren und könnte für die Äußerung keine vollständige Interpretation bereitstellen. Selbst vermeintlich unwichtige Äußerungspartikel wie „ehm“ können in besonderen Kontexten semantisch bedeutsam sein, z. B. können sie in spontansprachlichen Äußerungen einen Neuanfang oder eine Korrektur markieren [Kro00, Fis96]. Daher muss bei der Erstellung des Lexikons genau darauf geachtet werden, welche Wörter ignoriert werden können und es ist sinnvoll, zuerst so viele Wörter wie möglich aus dem Korpus aufzunehmen.

Homonyme können in dem Lexikon ebenfalls repräsentiert werden. Diese enthalten entsprechend ihrer Bedeutung Verweise auf unterschiedliche SSU und besitzen ggf. auch unterschiedliche syntaktische Einträge. Homonyme erhalten zur einfachen Handhabung jeweils ihren eigenen entsprechenden Lexikoneintrag, das heißt, zwei gleich lautende Wörter mit semantisch unterschiedlicher Bedeutung werden als zwei Lexikoneinträge realisiert. Das englische Wort „can“ beispielsweise (mögliche Übersetzungen sind „Dose“ oder „können“) kann mit dem Konzept *Container* oder auch *Fähigkeit_zum_Tun* (SSU *Ability*) verknüpft werden. Auch zu dem Konzept *Konservieren* kann das Wort eine mögliche Relation besitzen, jedoch ist es für unser *Hometour-Korpus* nicht relevant und wurde daher nicht in dem Lexikon aufgenommen. In anderen Kontexten kann das Lexikon bei Bedarf entsprechend adaptiert werden und diese Bedeutung mit aufführen. Kontextinformationen, die aus der gesamten Äußerung sowie dem Diskurswissen gewonnen werden können, geben Aufschluss, welches Homonym gemeint sein kann und daher für den Verarbeitungsprozess ausgewählt wird (siehe Kapitel 9). Weitere Homonyme sind z. B. die Aktion „clean“ und das Attribut „clean“, die Frucht „orange“ und die Farbe „orange“ oder das Element „water“ und die Aktion „water“.

Um einerseits das Lexikon und auch die semantischen Konzepte unabhängig vom System zu halten, und somit die Wissensbasen erweitern oder austauschen zu können und um andererseits leicht auf die Daten zugreifen zu können, wurden beide Datenstrukturen extern in einer Datei in Form von XML-Schemata gespeichert. Für unser System existiert jeweils ein Lexikon sowohl für ein deutsches als auch für ein englisches Korpus. Das deutsche Korpus enthält vorwiegend Wörter aus dem *Blumengieß-Szenario* (siehe Kapitel 6) und kam nur prototypisch zum Einsatz [Hüw05]. Das englische Lexikon basiert auf der Übersetzung des deutschen Lexikons und wurde zusätzlich um Wörter aus dem *Hometour-Korpus* erweitert. Ein weiteres Lexikon für das Deutsche existiert ebenfalls für das *Hometour-Korpus*. Es ist geplant, dieses Lexikon ebenfalls für die Interaktion mit BIRON einzusetzen.[1] Bei der Erstellung der Lexika wurde der gesamte Wortschatz der Korpora der *Hometour* und *Blumengießen* abgebildet und jeder Lexikoneintrag referenziert auf ein semantisches Konzept. Aufgrund neu entstandener Anforderungen, die durch Funktionserweiterungen des Robotersystems verursacht werden, wie z. B. autonome Navigation sowie der Erweiterungen des Sprachumfangs mit Hilfe von aufgenommenen Experimentdaten, befindet sich das Lexikon in einem kontinuierlichen Wachstumsprozess. Mittlerweile enthält das englische Lexikon etwa 1400 Einträge.

Im Folgenden werden einige Beispiele aus dem Lexikon dargestellt. Diese Beispiele enthalten Auszüge aus dem Lexikon des *Hometour-Korpus* in englischer Sprache. Ein großer Bereich des Lexikons besteht aus für die Interaktion mit dem Roboter wichtigen *Objekten*. Darunter fallen Wohnungsgegenstände, aber auch Lebensmittel oder Räume. Die Lexikoneinträge „Tisch“ und „Wohnzimmer“ in Abbildung 8.1 gehören zur Klasse der *Objekte*. Diese Information ist hier jedoch nicht explizit abgelegt, sondern implizit über die zugehörige SSU kodiert. Da sie in diesem Kontext für den Roboter unterschiedliche Funktionen besitzen, verweisen sie auf jeweils unterschiedliche SSUs, die jedoch alle zur Klasse der Objekte gehören (siehe Kap. 8.2.3). Die Lexi-

[1] In den Studien soll die Interaktion mit muttersprachlichen Probanden verglichen werden mit den Studien, die auf dem englischen Korpus mit Nicht-Muttersprachlern durchgeführt wurden.

```
<entry name="table">
   <SSU>Object_home</SSU>
   <WordType>n</WordType>
   <Syn>
      <Numerus>sing</Numerus>
   </Syn>
</entry>
```

```
<entry name="living_room">
   <SSU>Building_subpart</SSU>
   <WordType>n</WordType>
   <Syn>
      <Numerus>sing</Numerus>
   </Syn>
</entry>
```

Abbildung 8.1.: Lexikoneinträge verschiedener Objekte

koneinträge enthalten jeweils einen Namen und die zugehörige SSU. Zusätzlich werden wichtige syntaktische Informationen festgehalten, wie beispielsweise die Zuordnung der entsprechenden Wortart *Nomen*, abgekürzt mit „n“. Unter dem *Tag* „Syn“ des XML-Schemas ist noch zusätzlich die Information über den Numerus gespeichert. Im deutschen Lexikon sind darunter noch der Kasus und Genus gespeichert, da diese Informationen bei der Auflösung von Anaphern hilfreich sein können.

Die Aktionen „follow“ und „clean“ sind in Abbildung 8.2 dargestellt. Sie werden durch die verschiedenen SSUs *Move_afterwards* und *Cleaning* semantisch repräsentiert. Zusätzlich enthalten die Lexikoneinträge die gleichen syntaktischen Informationen. Die Verben, abgekürzt mit „v“, stehen in der 1. und 2. Person im Singular und in der 1., 2. und 3. Person im Plural, jeweils im Präsens.

```
<entry name="follow">
   <SSU>Move_afterwards</SSU>
   <WordType>v</WordType>
   <Syn>
      <Person>1|2</Person>
      <Tempus>present</Tempus>
      <Numerus>sing</Numerus>
   </Syn>
   <Syn>
      <Person>1|2|3</Person>
      <Tempus>present</Tempus>
      <Numerus>plur</Numerus>
   </Syn>
</entry>
```

```
<entry name="clean">
   <SSU>Cleaning/SSU>
   <WordType>v</WordType>
   <Syn>
      <Person>1|2</Person>
      <Tempus>present</Tempus>
      <Numerus>sing</Numerus>
   </Syn>
   <Syn>
      <Person>1|2|3</Person>
      <Tempus>present</Tempus>
      <Numerus>plur</Numerus>
   </Syn>
</entry>
```

Abbildung 8.2.: Lexikoneinträge von verschiedenen Aktionen

Homonyme enthalten entsprechende Verweise auf die in Relation stehenden semantischen Konzepte und ggf. auch auf unterschiedliche syntaktische Einträge. Der Behälter „can“ (*Container*) und das Modalverb „can“ (*Ability*), in Abbildung 8.3 dargestellt, zeigen die verschiedenen Einträge der Homonyme. Bei diesen Einträgen unterscheiden sich ebenfalls die syntaktischen Informationen deutlich voneinander. Der Lexikoneintrag „that“ (siehe Abb. 8.4) kann ebenfalls auf

zwei Arten interpretiert werden. Einerseits kann es einen Hinweis auf eine Geste des Interaktionspartners geben (*Maybe_gesture*), andererseits kann es anaphorisch verwendet worden sein (*Object_anaphoric*) und die fehlende Information kann allein aus vorhergegangenen Äußerungswissen gewonnen werden. Im Gegensatz zu „those“ steht „that“ im Singular.

```
<entry name="can">
   <SSU>Ability</SSU>
   <WordType>aux</WordType>
   <Syn>
      <Person>1|2|3</Person>
      <Tempus>present</Tempus>
      <Numerus>sing|plur</Numerus>
   </Syn>
</entry>
```

```
<entry name="can">
   <SSU>Container</SSU>
   <WordType>n</WordType>
   <Syn>
      <Numerus>sing</Numerus>
   </Syn>
</entry>
```

Abbildung 8.3.: Lexikoneinträge der Homonyme „can“.

```
<entry name="that">
   <SSU>Maybe_gesture</SSU>
   <WordType>det</WordType>
   <Syn>
      <Numerus>sing</Numerus>
   </Syn>
</entry>
```

```
<entry name="that">
   <SSU>Object_anaphoric</SSU>
   <WordType>det</WordType>
   <Syn>
      <Numerus>sing</Numerus>
   </Syn>
</entry>
```

Abbildung 8.4.: Lexikoneinträge des Wortes „that“

Weitere wichtige Konzepte für die Interaktion mit einem Roboter sind Fragenkonzepte. Auch hier werden unterschiedliche Realisierungen umgesetzt. Die Lexikoneinträge „what“ und „who“ im Englischen beispielsweise werden durch die SSUs *Question_action* oder *Question_person* repräsentiert. Der Eintrag „what“ oder im Deutschen „was“ hat ebenfalls mehrere Einträge, es kann auch nach Objekten, Personen, Attributen, usw. gefragt werden. Generell könnte der Eintrag auf das allgemeine Konzept *Question* verweisen. Im *Hometour-Kontext* jedoch wurde sich für die feineren Unterscheidungen entschieden, da so der Dialogmanager direkt durch die SSU erkennen kann, wonach gefragt wurde.

```
<entry name="what">
   <SSU>Question_action</SSU>
   <WordType>prep</WordType>
   <Syn>
   </Syn>
</entry>
```

```
<entry name="who">
   <SSU>Question_person/SSU>
   <WordType>prep</WordType>
   <Syn>
   </Syn>
</entry>
```

Abbildung 8.5.: Lexikoneinträge von Fragewörtern

8.2. Situierte semantische Konzepte

Neben dem Lexikon ist das Wissen über die Verwendung der Wörter, die semantische und pragmatische Bedeutung von vollständigen Äußerungen wichtig, damit der Roboter mit dem Benutzer interagieren kann. Dabei fließt die Besonderheit situierter Kommunikation stark in die Art der Wissensrepräsentation ein. Gängige Verfahren wie FrameNet [Bak98] sind nicht ausgelegt, Interaktionen zwischen zwei Kommunikationspartnern in einer realen Umgebung zu verarbeiten, denn das Korpus von *FrameNet* besteht aus grammatikalisch korrekten Sätzen aus dem Gebiet der Nachrichten und ist nicht für die automatische Sprachverarbeitung ausgelegt. Das spezielle Ziel bei FrameNet ist die Erstellung einer semantischen Datenbank.

Auch wenn in vielen Bereichen bereits Konzepte für die Repräsentation semantischer Informationen erstellt wurden, wird keine der uns bekannten Repräsentationsformalismen unseren Anforderungen für die Domäne der *Robot Companion* gerecht (siehe auch Kap. 3.2). Daher wurde ein eigener Repräsentationsformalismus entwickelt, der es ermöglicht, alle semantisch verwertbaren Äußerungen der Mensch-Roboter-Kommunikation analysieren zu können und so den Dialog durch sinnvolle semantische Interpretation zu unterstützen.

Das zentrale Anliegen ist, das benötigte Wissen der Domäne möglichst vollständig zu repräsentieren. Jedes Wort aus dem Korpus wird auf ein semantisches Konzept abgebildet. Diese semantischen Konzepte wiederum stehen in Relationen zueinander, sie sind sozusagen semantisch verlinkt. Dieses Konzept ermöglicht automatische Sprachverarbeitung – nicht nach sonst üblichen syntaktischen Merkmalen – sondern auf Basis der semantischen Beziehungen zueinander. Gerade für die Verarbeitung von situierter Spontansprache können diese Merkmale der Strukturen äußerst hilfreich sein. Zur Zeit sind etwa 150 semantische Konzepte in Form von SSUs realisiert worden, die die relevanten semantischen Informationen für das Korpus der Mensch-Roboter-Interaktion abbilden.

8.2.1. Grundidee

Die Idee der *situierten semantischen Einheiten* – der SSUs – basiert auf zwei Grundgedanken. Zum einen lehnt sie an die Theta-Rollen-Theorie [Fil68] an (siehe Kap. 3.1.2). Es wird angenommen, dass die Bestandteile eines Satzes oder einer Äußerung mit Hilfe so genannter Rollen beschrieben werden können. So können bestimmte Komponenten in einem Satz auf bestimmte Funktionen oder Rollen abgebildet werden. Zum anderen wird davon ausgegangen, dass die semantischen Merkmale einer Äußerung in eine Art semantisches Netzwerk überführt werden können. Ähnlich wie in der Konzeptuellen Semantik [Jac83, Jac90] wird angenommen, dass sich die Bedeutung eines Satzes in der Regel aus der Komposition der Wortbedeutungen erschließen läßt: Die Bedeutung eines Satzes ist eine konzeptuelle Struktur. Im Gegensatz zur konzeptuellen Semantik wird in diesem Ansatz nicht die Äußerung auf eine geringe Menge primitiver Konzepte reduziert, sondern es wird von den im Kontext der Interaktion benötigten Kategorien und deren notwendigen Unterscheidungen in der Kommunikationssituation ausgegangen. Die-

se Unterscheidungen der Kategorien wird demnach vom Bedarf gesteuert und unterliegt keinen theoretischen Modellen.[2]

Der zentrale Kerngedanke bei dem Konzept der SSUs ist, dass die semantische Bedeutung eines linguistischen Ausdrucks durch die Verlinkung von Konzepteinheiten – hier SSUs – repräsentiert werden kann. Dabei führen bestimmte Konzepte weitere Konzepte ein, sie triggern sozusagen weitere semantische Informationen und bilden die Wurzel des Netzwerkes. Selbst wenn diese Informationen nicht explizit genannt werden, kann sich der Mensch ein mentales Bild darüber machen. Das Wort „laufen“ beispielsweise triggert weitere beteiligte Konzepte. Es kann auf die SSU *Bewegung* abgebildet werden, die ein *Akteur* an einem bestimmten *Ort* oder auf ein bestimmtes *Ziel* hin ausführt. Ebenso wie eine Handlung kann ein Gegenstand auf weitere Konzepte verweisen. Ein „Tisch“ ist beispielsweise ein *Objekt* mit besonderen Eigenschaften und Merkmalen, es hat eine bestimmte *Größe*, steht an einem bestimmten *Ort* und hat eine bestimmte *Farbe*.

Unter der Annahme, dass jede SSU auf die in Relation zu ihr stehenden SSUs verweist, lassen sich semantische Templates erzeugen, die genau diese Relationen abbilden. Es wird somit ein Netzwerk semantischer Beziehungen geschaffen, in dem die SSUs untereinander semantische Relationen abbilden. Dadurch entsteht die Möglichkeit, einzelne Wörter und ganze Äußerungen miteinander in Beziehung zu setzen und aus einzelnen Einheiten eine kohärente semantische Struktur zu gewinnen. Das Konzept der Wissensrepräsentation bildet somit die Grundlage für die automatische Gewinnung semantischer Interpretationen von linguistischen Ausdrücken.

Dabei wird in diesem Ansatz die Zuordnung der Relationen der SSUs untereinander nicht aufgrund syntaktischer Eigenschaften, sondern auf rein semantischer Ebene vorgenommen. Entgegen dem Ansatz der Kasusgrammatik oder Dependenzgrammatik (siehe Kap. 3.3) bestimmt nicht die Wortart, insbesondere das Verb, ob und wie die SSU in Relation zu anderen SSUs steht, sondern einzig und allein ihre semantische Klassifizierung. Gerade in der Spontansprache können die syntaktischen Klassifikationen vom konkreten Gebrauch abweichen. Beispielsweise kann die Äußerung „Drehung rechts“ dasselbe bedeuten wie „dreh dich nach rechts“, wobei im ersten Fall die Aktion wesentlich durch das Nomen „Drehung“ bestimmt wird und im zweiten Fall die Aktion durch das Verb „drehen“. Daher werden auch Nomen wie „Drehung“ auf ein Aktionskonzept abgebildet, in diesem Fall auf der SSU *Move_in_place*. Die semantische Bedeutung bestimmt demzufolge die zugehörige SSU und nicht die Wortart.

8.2.2. Aufbau der situierten semantischen Einheiten

In der Datenbank der SSUs werden alle für die Interaktion wichtigen semantischen Informationen in Form von XML-Schemata gespeichert. Die SSUs wiederum bestehen aus einer Reihe von Informationen, die jeweils in den XML-Entitäten gespeichert sind. Die wichtigste Informa-

[2] Die Kategorien oder Konzepte sind auf der Basis von Korpora erstellt worden, die im Rahmen von Mensch-Roboter-Experimenten entstanden sind. Es wird kein Anspruch auf vollständige Korrektheit erhoben, es zeigte sich jedoch die Praktikabilität im laufenden System.

tion jeder SSU ist der Name, der als erstes Element gespeichert ist. Dieser Name ist eindeutig und dient der Diskriminierung der unterschiedlichen Konzepte. In Gegensatz zum Lexikon kann immer nur genau eine semantische SSU unter gleichem Namen auftreten.

Die Vernetzung der SSUs untereinander entsteht durch die semantische Zugehörigkeit einzelner SSUs zu anderen SSUs. Sie werden quasi semantisch getriggert, so wie ein „Wort“ oder Konzept jeweils bestimmte Assoziationen zu anderen „Wörtern“ weckt. Dabei sind nicht alle Assoziationen gleich stark. In den SSUs wird daher zwischen starken und schwachen Relationen unterschieden. Diese unterschiedlichen Arten der Verknüpfungen werden durch die verschiedenen Elemente der notwendigen (*mandatory*) und optionalen (*optional*) SSU-Relationen markiert.

Diese Unterscheidung wird getroffen, um differenzieren zu können, welche Informationen notwendig für das Verständnis sind, und welche als Ergänzungen dienen. Es werden genau dann starke Verknüpfungen verwendet, wenn der Mensch Probleme hätte, den Sinn ohne diese Informationen zu verstehen. Die optionalen Informationen sind für das grundsätzliche Verständnis des Sachverhaltes nicht wichtig, können aber hilfreiche Zusatzinformationen liefern. Die Kernaussage einer Äußerung wird durch das Weglassen der schwachen Verknüpfungen nicht beeinträchtigt. Fehlen jedoch Informationen aus starken Verknüpfungen, kann der Dialog nicht abgeschlossen werden und eine Rückfrage ist notwendig, um den Sinn der Äußerung zu entnehmen. Auch in vielen Grammatiktheorien findet eine Unterscheidung zwischen Kernbestandteilen und optionalen Argumenten statt. Beispielsweise wird in Phrasenstrukturgrammatiken zwischen Komplementen und Adjunkten unterschieden (vgl. [Bor91]). In dem hier vorgestellten Ansatz sind grammatikalische Eigenschaften jedoch nicht das tragende Unterscheidungsmerkmal, sondern die semantischen Beziehungen der einzelnen Bestandteile der Äußerung untereinander.

Die Unterscheidungsmöglichkeit bietet wichtige Stützen für die Interpretation der Äußerungen und für die Bereitstellung der Informationen. Mit der Möglichkeit, die Informationen, die mit einer SSU verknüpft sind, nach Wichtigkeit unterscheiden zu können, kann die Bewertung der Äußerungsinterpretation unterstützt werden. Sind in einer Äußerung alle notwendigen Informationen vorhanden, ist die Äußerung semantisch kohärent und die Äußerung ergibt einen Sinn. Fehlen jedoch zentrale Informationen, so kann die Bedeutung nur teilweise ausgemacht werden. Wenn nur optionale Informationen fehlen, kann dennoch der Sinn der Äußerung erschlossen werden. Anhand der Information, wieviele notwendige Argumente fehlen, kann die erkannte Äußerung bewertet werden. Dieses Kriterium gibt daher Aufschluss darüber, ob der Dialogmanager Rückfragen stellen muss oder ob der Dialog soweit abgeschlossen ist.

Darüberhinaus bietet dieses Unterscheidungsmerkmal der notwendigen und optionalen Elemente Hilfen für das Wissen über den konkreten Inhalt, den der Dialogmanager zum vollständigen Verständnis der Äußerung benötigt. Die mit den in der Äußerung in Beziehung stehenden SSUs liefern die Platzhalter der notwendigen Informationen für das Verständnis. Fehlen Teile, so kann zumindest weitergegeben werden, welche Informationen in Beziehung mit den SSUs stehen und welche bei der Analyse (siehe Kap. 9) gefüllt werden konnten und welche leer geblieben sind. Die Platzhalter der SSUs geben Hinweise, welche Information konkret fehlen. Diese Lücken können dann auf verschiedenen Wegen gefüllt werden. Zum Teil können die Informationen gezielt durch zusätzliches Szenewissen oder durch die Historie der Äußerungen gewonnen werden.

Der Dialogmanager kann ebenfalls gezielte Nachfragen an die Interaktionspartner stellen, um das benötigte Wissen zu erhalten.

Das letzte Element der XML-Struktur einer SSU ist die Angabe der hierarchischen Zuordnung. Diese Information wird in dem Element *top* gespeichert. Die Idee der hierarchischen Struktur wird im nächsten Abschnitt ausführlicher erläutert.

Abbildung 8.6 zeigt eine vereinfachte Darstellung der SSU *Showing*. Die notwendigen Elemente der SSU sind die in Verbindung stehenden SSUs *Actor* und *Object*. Optionale Elemente sind die SSUs *Person_involved* und *Time*. Die SSU *Showing* gehört im *Hometour*-Szenario zur Klasse *Task*.

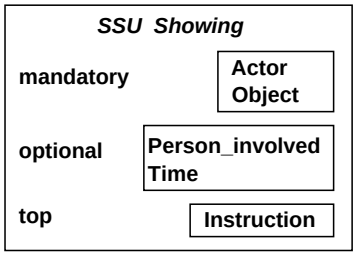


Abbildung 8.6.: Eine vereinfachte Darstellung der SSU *Showing*

8.2.3. Klassifikation und Hierarchiekonzept

Für die Repräsentation semantischer Zusammenhänge und semantischer Unterschiede werden spezifische SSUs benötigt. Eine Banane gehört zur Klasse der Lebensmittel (SSU *Food*), ein Radio ist ein gebräuchlicher Wohnungsgegenstand (SSU *Objekt_home*) und eine Tür ist Bestandteil eines Gebäudes (SSU *Building_subpart*). Diese Klassen können jeweils unterschiedliche Eigenschaften besitzen sowie auch die verschiedenen SSUs die für sie relevanten Informationen speichern können. Im System wird die Zuordnung spezifischer Informationen u. a. dadurch realisiert, dass die SSUs jeweils wiederum auf unterschiedliche SSUs referenzieren. Beispielsweise wird ein Gegenstand wie ein Tisch oder ein Radio in der Regel mit Hilfe seiner Farbe, seiner Größe oder seinem Standort beschrieben. Im Unterschied zu einem Radio oder anderen Gegenständen in der Wohnung besitzen Lebensmittel auch die Eigenschaft eines Geschmacks oder Aroma. Daher enthält die SSU *Food* im Gegensatz zur SSU *Objekt_home* (siehe Abb. 8.7) neben den Referenzen auf die SSUs wie *Position*, *Dimension* oder *Color*, eine zusätzliche Verknüpfung zur SSU *Flavor*. Die Zuordnung der SSUs und die feinere Unterteilung wurde vor allem nach relevanten Gesichtspunkten für den Einsatz von mobilen Robotersystemen erstellt (vgl. 8.2.4). Eine „Tür“ stellt ein *Building_subpart* dar, es ist wichtig für die räumliche Wahrnehmung und Navigation des Robotersystems. Es könnte in einem anderen Szenario jedoch auch eine Art „bewegliches Klappteil“ darstellen, das man bewegen kann. Das wäre z. B. für ein spezielles Roboter-Szenario interessant, in dem der Roboter Manipulatoren besitzt, mit denen er Türen öffnen kann.

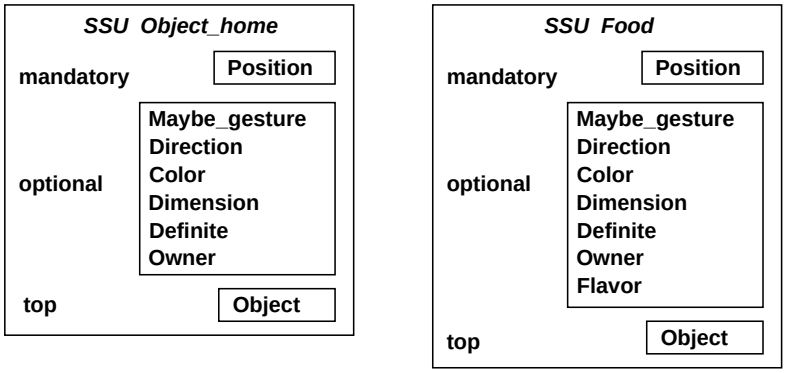


Abbildung 8.7.: Darstellung der SSUs *Object_home* und *Food*

Genauso wie man syntaktisch wohlgeformte Äußerungen von syntaktisch falschen Äußerungen unterscheiden kann, so gilt dasselbe auch für die semantischen Beziehungen der Äußerungsbestandteile untereinander[3]. Dass die folgenden Aussagen semantisch einen Sinn ergeben, ist nicht gleich wahrscheinlich:

(1) Er isst einen Apfel.

(2) Er isst einen Tisch. *[4]

Es existieren daher Konzepte (in diesen Beispielen für „essen“), die auf spezielle SSUs referenzieren. Anderseits gibt es semantische Konzepte wie *Showing* oder *Giving*, die auf eine Vielzahl von SSUs referenzieren können. Um nicht im SSU *Showing* alle möglichen SSUs auflisten zu müssen, die „gezeigt“ werden können, dennoch aber die semantischen Feinheiten der verschiedenen SSUs darstellen zu können, wurde eine hierarchische Abstraktionsstruktur eingeführt. Die situierten semantischen Einheiten (SSUs) sind daher durch so genannte Basis- oder Oberkategorien gruppiert. Ohne eine solche Struktur wäre eine Erweiterung des Lexikons mit zunehmendem Umfang immer komplexer. Die zugehörige Oberklasse ist in dem *top* Element der jeweiligen SSU vermerkt. Dadurch können die SSUs nicht nur direkt auf andere SSUs referenzieren, sondern ebenfalls auf die Oberkategorien. Das bedeutet, die SSU-Verknüpfungen können somit so generell wie möglich beschrieben werden. Gleichzeitig können semantische Feinheiten durch die Bereitstellung spezifischer SSUs abgebildet werden.

Die SSUs wurden in möglichst klar verständliche und überschaubare Kategorien eingeteilt. Zum einen soll die Strukturierung der Daten möglichst verständlich und dadurch leicht zu erweitern sein und zum anderen eine für die Verarbeitung möglichst sinnvolle Strukturierung darstellen. Im wesentlichen existieren die Oberkategorien *Task, Object, Actor, Description, Attribute, Position, Question, Socialisation*.

[3] Hier muss jedoch immer auch die aktuelle Situation berücksichtigt werden.
[4] Semantisch fragwürdig, es sei denn „er“ bezieht sich auf einen Holzwurm.

Nicht alle SSUs sind in diese Klassifizierung eingeteilt und besitzen daher eine Top-Kategorie. Das sind u. a. *Calendric_unit* wie „Montag“ oder *Cardinal_number* wie „drei“. Bei diesen SSUs ist eine Gruppierung in weitere Oberkategorien nicht sinnvoll, sondern würde den Aufbau nur unnötig komplexer machen. Einige Lexikoneinträge sind direkt einer Oberkategorie zugeordnet, da eine genauere Beschreibung für die Kommunikationsfähigkeiten nicht notwendig war. Für sie existieren keine weiteren SSUs. Beispielsweise sind nicht alle Objekte in weitere SSUs unterteilt, sie gehören direkt zur SSU *Object*, die gleichzeitig die Oberklasse der Objekte darstellt.

In der Regel werden die SSUs nur auf zwei Hierarchieebenen abgebildet. Dennoch existiert theoretisch keine Beschränkung. Einige SSUs bilden auch mehrere Ebenen ab. Das SSU *Name_personal* z. B. gehört zur Klasse *Actor*. Die Oberkategorie von *Actor* wiederum ist *Object*. Die Oberklasse *Position* gehört wiederum zur Klasse der *Attribute*. In den Tabellen 8.1 bis 8.8 werden die Oberkategorien und wichtige zugehörige SSUs aufgeführt. Um ein besseres Verständnis für die SSUs zu gewinnen, werden exemplarisch Lexikoneinträge angegeben, die auf die jeweilige SSU verweisen.

SSU-Name	**zugeortnete Wörter**
Object	„tool“, „toy“
Building_subpart	„bathroom“, „door“, „window“
Container	„bin“, „box“, „can“
Food	„apple“, „biscuits“
Object_soil	„coffee“ „milk“, „tea“
Object_office	„paper“, „pen“
Object_kitchen	„baking_dish“, „microwave“
Object_home	„bed“, „book“, „chair“
Object_anaphoric	„it“, „that“

Tabelle 8.1.: Tabelle der Klasse der *Objekte*

Die größte Klasse der SSUs ist die der *Objekte*. Sie enthält Objekte, über die kommuniziert werden kann. Das sind Gegenstände, die in der Szene zu sehen sind, die aber auch im Gedächtnis des Roboters gespeichert sein können. Die SSUs innerhalb der Klasse enthalten zum einen die spezifischen Informationen, welche semantischen Beziehungen sie zu anderen SSUs besitzen, also welche Verknüpfungen zu anderen SSUs existieren. Andererseits wurden Unterscheidungskriterien, die für das Robotersystem relevant sind, berücksichtigt (siehe Kap. 8.2.4).

Die Klasse der *Tasks* oder auch Handlungen enthält wichtige Aufgaben, die der Roboter ausführen können soll oder die an den Roboter gestellt werden können. Grundlegende Aktionen, die für die Mobilität von mobilen Robotern relevant sind, sind Bewegungsaktionen sowie das Nehmen und Tragen von Objekten. Um neue Gegenstände lernen zu können sind Handlungen wie Sprechen, Zeigen oder visuelles Wahrnehmen wichtig. Die Klasse *Description* umfasst die Menge von SSUs, die für die Beschreibung von Objekten, Personen oder Situationen verwendet werden. Hiermit können Äußerungen wie „das ist eine Banane“, „die Tasse ist grün“ oder „das klingt gut“ beschrieben werden. Die Äußerungen wie „das ist eine blaue Tasse “ und „ die Tasse

SSU-Name	zugeordnete Wörter
Task	
Carrying	„carry“
Change_topic	„talk“, „talking“, „change“
Move_in_place	„reverse“, „rotate“, „rotation“, „turn“
Move_afterwards	„come“, „follow“
Showing	„show“,
Looking	„check_up“, „look“, „look_after“, „see“
Take	„fetch“, „take“, „get“
Cleaning	„clean“, „wash“
Filling	„add“, „fill“, „refill“

Tabelle 8.2.: Tabelle der Klasse *Task*

ist blau“ könnten theoretisch ebenfalls durch unterschiedliche SSUs beschrieben werden. Hier hat sich jedoch im laufenden Betrieb gezeigt, dass diese Unterscheidungen für das Robotersystem nicht hilfreich sind. Daher wurden beide Varianten auf eine SSU zurückgeführt.

SSU-Name	zugeordnete Wörter
Description	
Existence	„are“, „am“, „was“, „were“
Name_bearing	„am“, „am_called“, „are_called“, „is“
Sound	„sound“, „sounds“

Tabelle 8.3.: Tabelle der Klasse *Description*

Die Klasse *Actor* bildet ebenfalls eine wichtige Kategorie für das Sprachverstehen. Hierbei wird deutlich, dass Syntax und Semantik einen Einfluss aufeinander haben. Die semantische Rolle eines „Besitzers“ (*Owner*) wird ebenfalls durch die syntaktische Form (Possessivpronomen) sichtbar. Dennoch ist bei den SSUs explizit nicht die syntaktische Form Unterscheidungsmerkmal, sondern die semantische Bedeutung ausschlaggebendes Kriterium. Die SSU *Owner* wurde in die Oberklasse *Actor* mit aufgenommen, auch wenn die Zuordnung nicht ganz unstrittig ist. In der Praxis hat sich diese Zuordnung jedoch als praktikabel erwiesen. Die Klasse der Aktoren selbst wiederum gehört zur Klasse der Objekte. Der Lexikoneintrag „roboter“ verweist auf eine SSU *Actor*, die wiederum ein *top* Element *Object* besitzt.

SSU-Name	zugeordnete Wörter
Actor	„robot“
Owner	„mine“, „my“, „our“, „your“
Proxy_personal	„he“, „I“, „it“, „she“, „they“, „we“, „you“
Proxy	„her“, „him“, „me“, „myself“, „us“, „you“
Name_personal	„biron“, „britta“, „jan“, „sonja“

Tabelle 8.4.: Tabelle der SSU-Klasse der *Aktoren*

Um Objekte oder Sachverhalte genauer beschreiben zu können, umfasst die Oberklasse *Attribute* die Menge der SSU, die bestimmte Eigenschaften beschreibt. Attribute sind beispielsweise die Farbe, die Farbintensität oder die Form. Die Oberklasse *Position* bildet hier eine Sonderform. Einerseits gehört die SSU *Position* zur Klasse der Attribute. Gleichzeitig bildet sie eine eigene Oberklasse. Unter dieser Oberklasse fallen die Präpositionen, die zur Beschreibung einer Position verwendet werden können. Diese Sonderform der Attribute ist gerade für die Situiertheit der Dialoge relevant, da Objekte in einem Raum vom System lokalisiert werden müssen.

SSU-Name	**zugeordnete Wörter**
Attribute	„clean“, „soft“
Color	„black“, „blue“, „lemon“, „orange“, „yellow“
Color_intensity	„bright“, „dark“, „light“
Dimension	„big“, „great“, „little“, „long“, „short“
Frequence	„always“, „daily“, „infrequent“, „weekly“

Tabelle 8.5.: Tabelle der SSU-Klasse der *Attribute*

SSU-Name	**zugeordnete Wörter**
Position	„aside“, „at“
Before	„ahead“, „before“ „in_front_of“
Down	„below“, „bottom“, „down“, „under“
Beside	„beside“, „close“, „left“, „near“
Behind	„after“, „behind“
In_between	„between“
On_top	„on“, „onto“

Tabelle 8.6.: Tabelle der SSU-Klasse der *Positionen*

Die in der Interaktion möglichen Fragen wurden in der Oberklasse *Question* zusammengefasst. Hier gibt die SSU Hinweise auf die Art der Frage, also nach welchen Informationen genau gefragt wird. Je nach Art der Frage, müssen vom System verschiedene Prozesse ablaufen - wird beispielsweise nach einem Objekt gefragt, tragen Informationen aus der Bildverarbeitung oder dem Gedächtnis zur Auflösung bei. Wird dagegen nach dem Thema gefragt (*Question_topic*), um das es im Dialog geht, werden Ergebnisse des Themendetektors benötigt. Durch die genaue Angabe der Fragestellung kann vom Dialogmanager direkt der richtige Prozess angesteuert werden und somit erhält das System einen schnellen Zugang zu den gesuchten Informationen. Die Fragekonstellation „Can you *xy*?“ ist etwas schwieriger zu klassifizieren. Hier handelt es sich in vielen Fällen um eine indirekte Aufforderung, bei der sich der Sinn der Äußerung in den meisten Fällen erst durch die Situation und Hintergrundwissen klärt. Hier wird daher das Wort „can“ als Hilfskonstrukt zu einer Anweisung gesehen (siehe Abschnitt 8.2.4). Auch kann es versteckte Fragen geben, wie „Du hast Geburtstag?“ Diese Aussagen können nur aufgrund zusätzlicher Prosodieinformationen erkannt werden. Das Erkennen von Fragen ist daher mitunter eine komplexe Aufgabe und nicht alleine mit Hilfe sprachlicher Informationen zu lösen.

SSU-Name	zugeordnete Wörter	mögliche Äußerungen
Question		
Question_action	„what“	„what can you do?“
Question_object	„what“, „which“	„what ist this? “, „which one? “
Question_attribute	„what“, „how“	„what color has *x*? “, „how big is the object? “
Question_name	„what“	„what is your name?“
Question_topic	„what“	„what are we talking about?“
Question_position	„where“, „what“	„where is *x*?“, „at what position is *x*?“
Question_person	„who“	„who is that?“
Question_time	„when“	„when do you leave?“

Tabelle 8.7.: Tabelle der SSU-Klasse der *Fragen*

Die Oberklasse der sozialen Interaktionen *Socialisation* spielt in der Interaktion mit einem *Robot Companion* ebenfalls eine Rolle. Wie aus dem Korpus ersichtlich, verwenden die Interaktionspartner entgegen der These des „Computer talks“ durchaus Höflichkeitsfloskeln. Soll der Roboter sich auf sein Gegenüber einstellen, so ist hilfreich, wenn diese sozialen Interaktionen auch verstanden werden können. Dann kann der Roboter entsprechend antworten. Auch zum Einstieg und zur Beendigung dient diese Klasse der SSUs.

SSU-Name	zugeordnete Wörter
Socialisation	
Greeting	„hello“, „hi“, „how“, „how_are_you“
Judgement_communication	„thank“, „thanks“, „thank_you“
Parting	„that’s_it“, „bye“, „bye-bye“, „good-bye“
Social_interaction	„please“

Tabelle 8.8.: Tabelle der Klasse der SSUs für Soziale Interaktion

In der Regel werden die SSUs nur auf zwei Hierarchieebenen abgebildet. Dennoch existiert keine Beschränkung. Einige SSUs bilden auch mehrere Ebenen ab. Die SSU *Name_personal* z. B. gehört zur Klasse *Actor*. Die Oberkategorie von *Actor* wiederum ist *Object*.

Die Hierarchisierung der SSUs bildet im Rahmen des Sprachverstehens einen sehr wichtigen Bestandteil im Gesamtkonzept. Ohne sie wäre eine effiziente automatische Verarbeitung nicht möglich (siehe Kap. 9). Für die Verarbeitung wird, wie bei der klassischen Unifikation, immer der allgemeinste Typ als so genanntes Muster genutzt. Daher benötigt das System Wissen über die hierarchischen Beziehungen der SSUs untereinander, um die Konzepte miteinander verbinden zu können und eine vollständige semantische Struktur gewinnen zu können. Durch die Hierarchiebeziehungen können die Beziehungen zueinander so generell wie möglich beschrieben werden. Dennoch können in den SSUs semantische Unterschiede festgehalten werden, die für die Kommunikationsfähigkeiten wichtig sein können oder eine Erleichterung für das Dialogsystem darstellen.

Zusätzlich lassen sich aus den Oberkategorien auch die Dialogakte ermitteln. Dabei können die meisten Kategorien eins-zu-eins in Dialogakte übertragen werden. Unter anderem sind das: *So-*

cialisation, *Task*, *Description* und *Question*. Da nicht alle SSUs einer Oberkategorie angehören, werden einige auch direkt auf die Dialogakte abgebildet. Darunter fallen *Confirmation*, *Negation*, *Deletion* und *Correction*. Der genaue Transformations-Prozess ist in Kapitel 9.6.3 beschrieben.

8.2.4. Situierte semantische Einheiten für den Roboter BIRON

Im Rahmen von *Robot Companions* werden besondere Konzepte für die erfolgreiche Interaktion benötigt. Der Roboter muss verstehen können, was die Interaktionspartner sagen, welche Erwartungen sie an ihn richten und wie er diese Erwartungen erfüllen kann. Er muss darüber hinaus auch Kontexte verstehen können, bei denen die Grenzen seiner Handlungsfähigkeiten überschritten sind. Nur dann kann er über diese Grenzen kommunizieren. Um Kommunikationsfähigkeiten über seine Aufgabenstellungen und seine Handlungsoptionen zu besitzen, benötigt das Robotersystem semantisches Wissen darüber. Auch auf Objekte in der Umgebung soll er mittels Sprache aufmerksam gemacht werden können. Ebenfalls soll er sich auf den Dialogstil seines Gegenübers einstellen können. Letztendlich geht es darum, Wissen über die aktuelle Kommunikationssituation zu erlangen und eine kommunikative Beziehung zu dem Gegenüber aufbauen zu können. Viele SSUs sind daher auf die Funktionalität des Robotersystems abgestimmt und dienen dem Verständnis der Anforderungen, denen das System gegenübersteht.

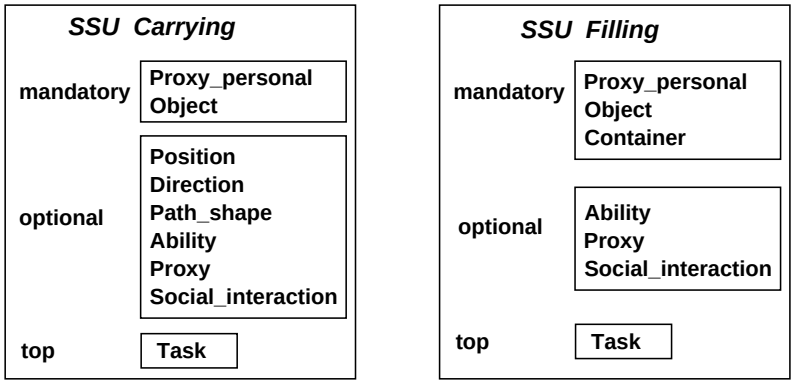


Abbildung 8.8.: Darstellung der SSUs *Carrying* und *Filling*

Ein wichtiger Bereich für *Robot Companions* ist die Übernahme von Aufgaben. Der Roboter soll den Benutzer mit seinen Fähigkeiten unterstützen. Damit ist die Darstellung verschiedener Aktionen verbunden. Für einen Haushaltsroboter sind z. B. die Aufgaben Reinigen, Gegenstände transportieren oder Blumengießen wichtige Handlungen, die kommuniziert werden müssen. Jedoch auch viele Grundfunktionen wie jemanden folgen, sich drehen können, etwas geben, etwas zeigen oder tragen, müssen dargestellt werden. Dafür stehen ebenfalls verschiedene SSUs zur Verfügung, die diese Aufgaben repräsentieren können, wie z. B. *Cleaning*, *Carrying*, *Filling*, *Giving*, *Looking*, *Move_afterwards*, *Move_in_place* oder *Showing*. In Tabelle 8.2 ist eine Übersicht wichtiger SSUs dargestellt, die Handlungen und Anweisungen abbilden. Die SSUs *Showing*, *Carrying* und *Filling* mit ihren Referenzen auf weitere SSUs sind in Abbildung 8.6 und 8.8 dargestellt.

Für das *Blumengieß-Szenario* existiert die SSU *Cause_to_be_wet* (siehe Abb. 8.9). Hiermit werden Wörter wie „gießen“ oder im Englischen „water“ repräsentiert.

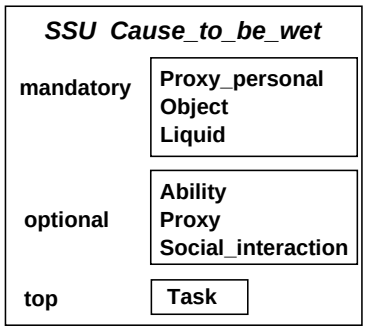


Abbildung 8.9.: Darstellung der SSU *Cause_to_be_wet*

Bei der Einführung des Roboters in die neue Umgebung im Rahmen der *Hometour* und auch bei der genauen Beschreibung der Aufgaben des Roboters ist es notwendig, dass die Umgebung und konkrete Objekte beschrieben werden können. Wenn der Roboter neue Objekte lernen können soll, sind die Informationen darüber ebenfalls relevant. Das System soll auch bisher unbekannte Personen kennen lernen und mit Namen ansprechen können. Objekte können neben dem Objektnamen durch Attribute genauer beschrieben werden oder in Relation zu anderen Objekten stehen. Für die Namensgebung wurde die SSU *Name_bearing* (siehe Abb. 8.10) definiert, das das Verb in „my name is Fridolin“ oder „this is BIRON“ repräsentiert. Das Wort „name“ wird auf die SSU *Naming* referenziert und das optionale Argument „this“ wird auf die SSU *Demonstrativ* referenziert. Für die Beschreibung von Objekten stellt die SSU *Existence* ein zentrales Konzept dar (siehe Abb. 8.10), das ebenfalls das Verb repräsentiert. Attribute und Objekteigenschaften werden durch eine Vielzahl von SSUs genauer definiert. Eine Auswahl wichtiger Attribute ist in den Tabellen 8.5 und 8.6 dargestellt.

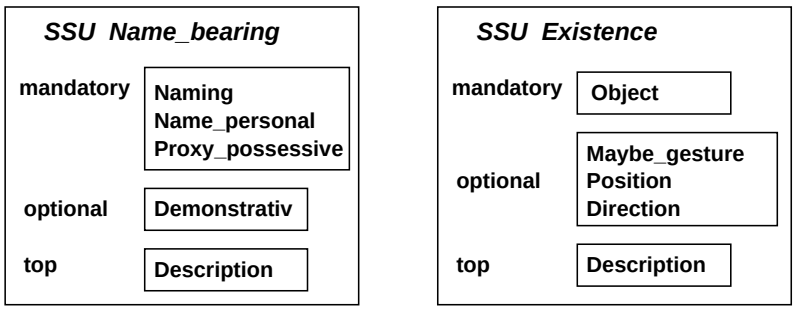


Abbildung 8.10.: Darstellung der SSUs *Name_bearing* und *Existence*

Die verschiedenen SSUs wurden für das Robotersystem nach relevanten Unterscheidungsmerkmalen definiert. Objekte werden daher nach wichtigen Kriterien und den konzeptuellen Zugehörigkeiten – wie bestimmte Handlungen oder Dialogsituationen – weiter unterteilt. Die SSU *Building_subpart* beispielsweise klassifiziert Objekte in einem Gebäude mit fester Position wie

„Tür“ oder „Fenster“. Der Roboter oder generell ein Akteur kann sich an diesen Objekten orientieren und seine Navigation danach ausrichten. Für das *Blumengieß-Szenario* existiert als Unterklasse der Objekte die Klasse der Pflanzen.

Zusätzlich zu den Kommunikationsfähigkeiten über Handlungen und Objekte sind die sozialen Aspekte ein wichtiger Bereich, um eine Beziehung zwischen Benutzer und Roboter herstellen zu können auf die eine gelungene Interaktion aufbaut. Damit der Roboter Aufgaben erledigen kann und die Kommunikation darüber überhaupt stattfindet, benötigt der Roboter Wissen über den sozialen Gebrauch der Sprache. Für den ersten Kontakt ist die Begrüßung wichtig, ebenso am Ende die Verabschiedung. Dafür werden jeweils unterschiedliche SSUs bereitgestellt, die diese sozialen Aspekte repräsentieren können. Um die generelle Kommunikationsfähigkeit des Roboters zu gewährleisten, sind eine Reihe weiterer SSUs von Bedeutung. Zustimmung, Rückfragen, Klärungen und die Verneinung sind wichtige Fähigkeiten, sowohl für den Menschen als auch für den Roboter. Dafür stehen die SSUs *Negation, Confirmation, Correction* und *Query* zur Verfügung. Diese bilden auch weitere Kategorien für die möglichen Dialogakte während der Interaktion.

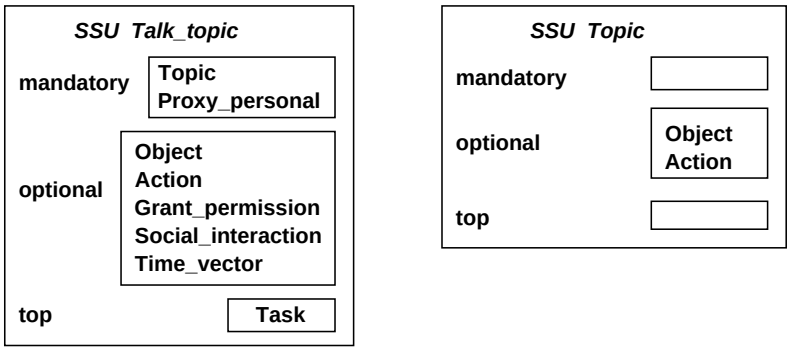


Abbildung 8.11.: Darstellung der SSUs *Talk_topic* und *Topic*

Eine besondere Gruppe von SSUs wurde aufgrund der Möglichkeit zur Themendetektion erstellt. Im Robotersystem BIRON ist ein Modul zur dynamischen Themendetektion integriert, das einen Themenwechsel neben impliziten Hinweisen zum Beispiel durch Ortswechsel, aber auch durch explizite Äußerungen (vgl. Kap. 5.8) erkennen kann. Dadurch kann der Roboter besser die aktuelle Situation überblicken und darin agieren (*situation awareness*). Im Bereich Themendetektion gibt es drei Kategorien, die sich in den SSUs widerspiegeln. Die erste ist die Bestimmung eines aktuellen Themas. Das kann z. B. durch die Äußerung „We talk about *x*.“ oder „Let us talk about *y*.“ angekündigt werden. Dafür wird die SSU *Talk_topic* genutzt, die in Abbildung 8.11 dargestellt ist. Das Thema kann ein Objekt oder eine Aktion beschreiben, wie z. B. „wir sprechen über die Küche“ oder „wir reden übers Abwaschen“. Das Wort „about“ oder „über“ wird auf die SSU *Topic* referenziert (siehe Abb. 8.11) und die beteiligte Person oder die beteiligten Personen auf die SSU *Proxy_personal*. Die SSU *Grant_permission* spiegelt das Wort „let“ wider und *Social_interaction* eine mögliche Höflichkeitsformel wie „bitte“. Die SSU *Time_vector* gibt zusätzlich Informationen über den genauen Zeitpunkt wieder.

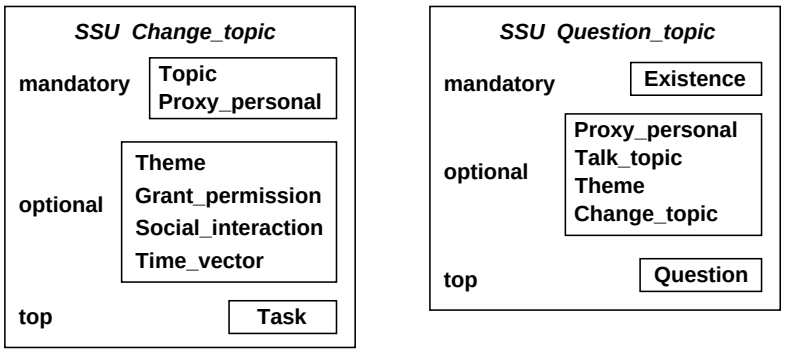


Abbildung 8.12.: Darstellung der SSUs *Change_topic* und Question_topic

Desweiteren können auch Themenwechsel angekündigt werden, z. B. durch die Äußerung „let us change the topic“. Hierfür dient die SSU *Change_topic* (siehe Abb. 8.12). Die SSU selbst verweist wiederum auf SSUs, z. B. der *Theme*, die Wörter wie „topic“ oder „theme“ abbilden.

Der letzte Bereich der Themendetektion sind Fragestellungen über das aktuelle Thema. Durch die SSU *Question_topic* (siehe Abb. 8.12) wird dem Benutzer die Möglichkeit gegeben, eine Anfrage an das Robotersystem zu stellen, er kann beispielsweise die Fragen „What are we talking about?“ oder „What is the topic?“ stellen.

8.2.5. Multimodalität

Die Dialoge zwischen dem mobilen Roboter und dem Interaktionspartner finden in einer realen Umgebung statt. Diese Situation entspricht der sonst üblichen Interaktionssituation zwischen zwei Menschen. So wie auch den Menschen in der Regel alle Kommunikationskanäle zur Verfügung stehen, so sollte für eine möglichst intuitive Interaktion mit dem Roboter das Robotersystem auch die Fähigkeit besitzen, neben der reinen Sprache zusätzlich Bildinformationen verarbeiten zu können. Darunter fallen vor allem die Gesten- und Objekterkennung aber mitunter auch die Verarbeitung von Mimik. Das Robotersystem benötigt ebenfalls die Möglichkeit, die sprachlichen Informationen mit den visuellen Informationen in Beziehung setzen zu können, um daraus zusammenhängendes Wissen generieren zu können, das für die Interaktion in der realen Welt notwendig ist.

Für die Interaktion mit dem Robotersystem BIRON stehen dem Benutzer daher mehrere Kommunikationskanäle zur Verfügung, die verarbeitet und vom System integriert werden müssen. Neben den rein sprachlichen Besonderheiten situierter Sprache werden daher auch die Hinweise in der Sprache auf visuelle Informationen durch das Sprachverstehen verarbeitet. Das System erhält mitunter Hinweise auf Zeigegesten durch bestimmte Äußerungskonstellationen. Die Äußerung „diese hier“ oder im englischen „look at this“ werden meistens zusammen mit einer Zeigegeste verwendet. Um die genaue Bedeutung aufzulösen, ist neben der Information der Äußerung die Auflösung der Referenz der Zeigegeste notwendig. Die Äußerung selbst kann zwar

nicht vollständig aufgelöst werden, sie bietet jedoch einen hilfreichen Hinweis auf sprachbegleitende Informationen. Wörter wie „diese“, oder im englischen „this“, „that“ oder „here“ können einen Hinweis auf vorhandene Gesten liefen. In der Wissensbasis werden daher die Hinweise aus der Sprache auf andere Modalitäten in den SSUs abgebildet. Die SSU *Maybe_gesture*, liefert daher für den Dialog eine wichtige Hilfe. Sie markiert, dass möglicherweise gestenbegleitende Informationen in der Szene vorhanden sind. Beispielsweise stellt sie für Objektbeschreibungen ein optionales Argument dar, wie in Abbildung 8.7 abgebildet.

Zusammen mit der genauen Zeitangabe der Äußerung kann die Geste, die auf das gesuchte Objekt referenziert, dann verarbeitet werden und mit dem Inhalt der Sprache in Zusammenhang gesetzt werden. Da Gestenerkennung ein zeit- und rechenintensives Verfahren ist und auch eher ungenaue Ergebnisse liefert [Haa05], kann der sprachliche Hinweis darauf das System wesentlich unterstützen und die Aufwandskosten stark reduzieren. Zudem erleichtert es die Integration der Multimodaltiät des Systems.

Wörter wie „diese“ können ebenfalls eine anaphorische Bedeutung besitzen. Leider ist nicht zu erkennen, ob sie eine Referenz auf die Szene ist oder ein Verweis auf bereits durch Sprache bekanntes Wissen. Nur wenn keine Geste verwendet wurde, trägt das entsprechende Wort eine anaphorische Bedeutung. Daher enthalten die Wörter, die ebenfalls anaphorisch verwendet werden können, entsprechend im Lexikon einen weiteren Eintrag, der auf diese mögliche Verwendung durch eine Referenz auf eine SSU *Anaphoric* verweist. Diese SSU markiert die anaphorische Verwendung und somit die Anaphernresolution für das System.

Einzig aufgrund der Sprachdaten kann das System zwischen den beiden Varianten nicht unterscheiden. Sowohl die SSU *Maybe_gesture* als auch die SSU *Anaphoric* markieren eine Referenz auf Wissen außerhalb der Äußerung. Ob das Wissen durch vorangegangene Äußerungen oder durch Szeneinformationen ergänzt werden kann, kann nicht allein aus den sprachverarbeitenden Prozessen gelöst werden. Nur wenn es keine Hinweise aus der Szene gibt, kann das System von einer anaphorischen Verwendung ausgehen. Aufgrund der Analyse der Experimente zwischen Roboter und Probanden und Überlegungen aus dem *Hometour*-Korpus wird angenommen, dass im Deutschen die Wörter „die“ und „das“ sowohl anaphorisch als auch gestenbegleitend verwendet werden können. Im Englischen sind das die Wörter „this“, „these“ und „that“. Das Wort „diese“ oder das englische Wort „this“ wird wie einige andere Wörter rein gestenbegleitend verwendet. Personalpronomen werden meist anaphorisch verwendet.

Die Verbindung der sprachlichen Informationen mit anderen Modalitäten wird in den SSUs ebenfalls durch die Bereitstellung von Referenzen auf andere SSUs verwirklicht. Objekte enthalten beispielsweise die Informationen *Dimension*, *Position* und *Color*. Diese Informationen können verbalisiert werden, aber auch durch die Informationen aus der Szene gewonnen werden. Ebenfalls kann das System in dem Szenemodell (siehe Kap. 5.4) nach diesen Informationen suchen. Die SSUs stellen daher auch eine Stütze dar, die fehlenden Informationen zu finden, um eine Situation oder ein Objekt vollständig zu erfassen und identifizieren zu können.

8.3. Diskussion

Die situierten semantischen Einheiten für die Mensch-Roboter-Kommunikation sind von Hand erstellt und gepflegt. Denkbar wäre jedoch auch die Definition und Entwicklung der situierten Konzepte im Rahmen eines automatischen Generierungstools wie in FrameNet [Fil01, Gil00]. Jedoch müsste man dafür eine große Datenmenge von Äußerungen zur Verfügung haben. Die Sammlung der Daten würde für sich einen enormen Aufwand bedeuten und wäre für das Szenario mit dem Roboter BIRON kaum leistbar. Hier war es wichtiger, dass die Datensammlung während der Entwicklung laufend durch Experimente erweitert werden konnte.

In dem Designprozess war grundsätzlich zu entscheiden, welche linguistische Einheit die semantische Wurzel einer Äußerung darstellt. Nicht immer war dies eindeutig und oftmals existierten alternative Lösungen. Die Auswahl der Wurzel und somit die Wahl der wurzelbildenden SSU war daher nicht immer leicht. Soweit möglich, wurde die Entscheidung aufgrund der Kategorisierung der Dialogakte vorgenommen. Der dahinter stehende Grundgedanke ist, den Dialogmanager so weit es geht zu unterstützen – und somit dem gesamten Kommunikationsprozess des Robotersystems zu erleichtern. Die SSUs wurden demnach weniger nach klassischen linguistischen Entscheidungen strukturiert, sondern nach den Kriterien der Relevanz für das Robotersystem, wobei sich jedoch auch viele Parallelen zeigten, insbesondere bei der Wahl der Regenten in der Dependenzgrammatik (vgl. [Mel88, Lob93a]).

Der erste Schritt bei der Entwicklung der SSUs stellte die Abbildung der zentralen Informationen dar, die aus der Korpussammlung gewonnen wurden. Aus dieser Basis heraus wurde ebenfalls die Abstraktionsebene der SSUs festgelegt und angepasst. Die Anforderungen an ein *Robot Companion* bildete die Entscheidungsgrundlage. Beispielsweise wurden *Objekte* nach wesentlichen Unterscheidungsmerkmalen definiert, *Aktionen* nach dem Handlungsrahmen eines Roboters. Dennoch lassen sich aufgrund der Hierarchisierung leicht Änderungen in dem Detailliertheitsgrad vornehmen. Auch die Verlinkung zu anderen Modalitäten spielte eine Rolle beim Desingkonzept der SSUs.

Neben den eher pragmatischen Designentscheidungen wurde auch theoretischeres Wissen integriert, insbesondere bei der Hierarchisierung. Folgendes Beispiel soll dieses verdeutlichen: Ein Akteur ist in der Regel eine Person, in Ausnahmefällen jedoch auch ein Objekt. Die Aussage „Der Regen lief mir in die Schuhe.“ im Kontrast zu „Die Nachbarin lief zum Baumarkt.“ ist eine gängige Formulierung, wenn auch der Regen (repräsentiert durch die SSU *Object_abstract*) nur im Übertragenen Sinn „läuft“. Genauso werden in der Regel „Pflanzen gegossen“, in bestimmten Kontexten kann es aber durchaus Sinn machen, andere Objekte bei der Interpretation zum Gieß-Vorgang mit einzubeziehen (z. B. Kerzen gießen). Deswegen ist für das Sprachverstehen der Äußerungen das Hierarchiekonzept der SSUs obligatorischer Bestandteil.

Für einen gelungenen Dialog sind eher die pragmatischen als die rein semantischen Interpretationen relevant. Daher werden die SSUs weitestgehend auf die pragmatischen Informationen abgebildet und die Äußerungen soweit möglich direkt in die pragmatische Bedeutung transformiert (z. B. indirekte Sprech-Akte wie „Kannst du mir die Butter geben?“). Redewendungen oder

Idiome stellen eine besondere Herausforderung für die Konzeption der SSUs dar. Hier kann die Bedeutung eines Satzes in der Regel nur im Ganzen und nicht aus seinen Teilen erschlossen werden. Diese Äußerungen besitzen neben der idiomatischen Bedeutung noch ihre wörtliche Bedeutung. Im kleineren Rahmen wird das Sprachverstehen auch mit dieser Art von Äußerungen konfrontiert. Diese Problematik kann dadurch gelöst werden, dass die Spracherkennung sowohl eine feste Redewendung als auch ein Ganzes an das Sprachverstehen weiterleitet. Diese Äußerung besteht dann sozusagen nur aus einem einzigen Lexikoneintrag und kann vom Sprachverstehen direkt in die übertragene Bedeutung überführt werden. Nicht immer gelingt die Transformation vom Erkenner direkt in ein Wort. Dann ist es mitunter aber ebenfalls möglich, die beiden Bedeutungen genauso zu behandeln wie Homonyme. Die Äußerungen „das wär‘s“ oder im Englischen „that’s it“ bedeuten in der Interaktion meistens eine Beendigung des Gespräches, zweiteres kann aber ebenfalls interpretiert werden als „du hast es erfasst“ oder eben wörtlich als „das ist es“. Für die englischen Dialoge besitzt das Wort „that“ oder im deutschen Lexikon das Wort „wäre“ ebenfalls eine Relation zur SSU *Parting*, die einen Abschied markiert. Entsprechendes gilt für die Äußerung „how are you“, die keine Frage darstellt, sondern eine Begrüßung.

Sowohl das Lexikon als auch die SSUs wurden im ersten Schritt für das deutsche *Blumengieß-Szenarios* erstellt. In einem zweiten Durchlauf wurde die gesamte Interaktion für den Roboter BIRON ins Englische übertragen. Ebenfalls wurde der Datensatz um das *Hometour-Korpus* erweitert. Für die Übersetzung ins Englische wurde das Lexikon direkt aus dem Deutschen übersetzt. Es stellte sich dabei heraus, dass es ausreicht, dafür die Wörter direkt zu übersetzen, und die syntaktischen Informationen zu ändern. Die Relationen zu den SSUs konnten direkt ohne Änderungen übernommen werden. Die SSUs selbst konnten bei der Übersetzung des *Blumengieß-Szenario* so bleiben – es waren keine Änderungen notwendig. Die SSUs mussten bei der Einbindung des *Hometour-Korpus* nur um neue Konzepte erweitert werden, ohne dass Änderungen notwendig waren. In einer erneuten Übertragung des englischen *Hometour-Korpus* ins Deutsche war wiederum bis auf eine Erweiterung die direkte Übernahme der SSUs möglich.[5] Das bedeutet, dass derselbe Datensatz von SSUs somit sowohl für das Englische als auch für das Deutsche eingesetzt werden kann. Selbst Mischformen von englisch und deutsch sind für das System ohne Probleme zu verstehen, so dass ein Wechsel innerhalb einer Dialogsequenz von Englisch nach Deutsch oder umgekehrt möglich ist[6]. Das Lexikon enthält dann entsprechend alle Wörter aus beiden Sprachen. Für andere Sprachen wurde die Übertragbarkeit nicht getestet, dennoch ist die Wahrscheinlichkeit groß, dass für verwandte Sprachen (z. B. Niederländisch) eine problemlose Erweiterung ebenfalls möglich ist.

Aufgrund der manuellen Erstellung der SSUs kann nicht garantiert werden, dass die Daten vollkommen korrekt, fehlerfrei und konsistent sind. So wie das Gesamtsystem evaluiert und getestet werden muss, müssen auch die Daten diesen Prozess durchlaufen. Dennoch zeigte sich im Einsatz, dass diese Methode einen guten Ansatz für das Sprachverstehen bietet (siehe Kap. 10). Die SSUs können daher als ein praktikabler Formalismus zur Wissensrepräsentation angesehen werden.

[5] Im Englischen wurde ein Ende mit „that’s_it“ vom Spracherkenner direkt als ein Wort weitergegeben, im Deutschen wurde der Datensatz dann um die Möglichkeit, eine vollständige Äußerung wie „(danke) das wär’s“ als End-Markierung zu akzeptieren, erweitert.

[6] unter der Annahme, dass die Spracherkennung für unterschiedliche Sprachen funktioniert

8.4. Zusammenfassung

Insgesamt stellen die SSUs ein Konzept für die Repräsentation von Welt- und Diskurswissen dar. Zusammenfassend lassen sich folgende Eigenschaften festhalten, die für das Verstehen von situierter Spontansprache eine wichtige Stütze darstellen:

- Das Lexikon enthält sowohl syntaktische als auch semantische Informationen.
- Die SSUs stellen ein Beschreibungsmittel für semantische Informationen dar.
- Im Lexikon können ebenfalls Homonyme dargestellt werden.
- Den etwa 1400 Lexikoneinträgen sind ca. 150 SSUs zugeordnet.
- Die SSUs sind für die automatische Sprachverarbeitung geeignet.
- Die Wissensbasen (Lexikon und SSUs) sind einfach zu erweitern und zu adaptieren.
- Vererbungsinformationen werden genutzt, um SSUs möglichst allgemein zu beschreiben und können zur Bereitstellung der Dialogakte genutzt werden.
- Die SSUs reflektieren Multimodalität: Sie stellen sprachliche Hinweise auf visuelle Szeneinformationen dar und sind zusätzlich geeignet, um visuelle Informationen darzustellen.
- Die SSUs liefern Hinweise auf die semantische Kohärenz und dadurch eine Infrastruktur für die Evaluation der Güte der Spracherkennungsleistung.
- Die SSUs sind für verwandte Sprachen wie Englisch und Deutsch weitestgehend sprachübergreifend.

9. Robuster Verarbeitungsprozess

Im vorangegangenen Kapitel wurden die Wissensbasen der sprachlichen Informationen für den Verstehensprozess von Äußerungen beschrieben. Dabei stellt sowohl lexikalisches als auch konzeptuelles Wissen wichtige Komponenten für den Interaktionsprozess des Robotersystems dar. Dieses Kapitel beschreibt den Verarbeitungsprozess, der diese Daten nutzt, um sprachliche Informationen eines Interaktionspartners zu einer semantischen Einheit zu konvertieren [Hüw06b]. Zusätzlich werden die vielfältigen Anforderungen (siehe Kapitel 7), die aufgrund der komplexen Interaktion zwischen Mensch und Roboter in einer realen Umgebung bestehen, vom Gesamtsystem berücksichtigt und in den Entwicklungsprozess der Sprachverstehenskomponente integriert. Kapitel 9.1 gibt zunächst einen Überblick über den Verarbeitungsmechanismus, der in Kapitel 9.2 detaillierter beschrieben wird. Die Besonderheiten des Ansatzes, Spontansprache zu verarbeiten, sowie die Behandlung von Anaphern werden in Kapitel 9.3 und 9.4 aufgezeigt. In Kapitel 9.5 wird zunächst die Testumgebung für den in dieser Arbeit vorgestellten Ansatz beschrieben und in Kapitel 9.6 wird schließlich die Integration auf die Roboterplattform BIRON erläutert.

9.1. Semantisches Parsing mit SSUs

Um semantische Interpretationen der Äußerungen an den mobilen Roboter zu generieren, wird ein semantischer Verarbeitungsmechanismus verwendet. Dieser erzeugt aus einer erkannten Wortkette eine möglichst vollständige semantische Interpretation oder Ableitungsstruktur. Wie auch beim syntaktischen Parsen werden mit diesem Ansatz verschiedene mögliche Ableitungen generiert. Bei der Verarbeitung wird aus einer Äußerung mit Hilfe des Lexikons und der konzeptuellen Einheiten – den *situierten semantischen Einheiten* (SSUs) – eine semantische Interpretation gewonnen. Diese Ableitung bildet sozusagen instanziierte SSUs ab.

Zunächst werden zu den Wörtern der Äußerung die entsprechenden Lexikoneinträge herausgefiltert. Existieren zu einzelnen Wörtern Homonyme (vgl. Kap. 8.1), werden diese ebenfalls entsprechend herausgesucht. Zu einem Wort existieren dann analog zu der Anzahl der Homonyme entsprechend viele Zuordnungen. Anschließend werden aufgrund der Verweise in den Lexikoneinträgen die in Beziehung stehenden SSUs zugeordnet. Innerhalb des Verarbeitungsmechanismus stellen die Wurzeln einer Äußerung oder Teiläußerung eine besondere Funktion dar. Sie entstehen implizit dadurch, dass auf sie keine andere SSU verweist und werden vom Mechanismus selbständig ermittelt. Einige SSUs sind potenzielle Wurzeln: in der Regel stellen Aktionen

oder Frage-SSUs eine Wurzel dar, in Ausnahmefällen können sie jedoch selbst auch von anderen SSUs verlinkt werden. Beispielsweise ist in der Aussage „was kochst du?“ die Aktion nicht die Wurzel, sondern das Fragewort „was“. Umgekehrt stellt in der Aussage „Petra fragte mich gestern, was ich koche.“ das Wort „fragte“ die Wurzel dar und nicht das Fragewort „was“. Das wiederum repräsentiert die Wurzel für die Teiläußerung „was ich koche“. Welche SSU also die Wurzel ist, hängt demnach im Wesentlichen von der semantischen Bedeutung ab. Das besondere an den SSUs, die eine potenzielle Wurzel einer Äußerug oder Teiläußerung darstellen, ist, dass sie offene Links zu anderen SSUs besitzen, die durch entsprechende SSUs, die andere Wörter bereitstellen, geschlossen werden. Dadurch ensteht ein Graph von miteinander verbundenen SSUs, die letztendlich einen semantischen Parsebaum ausbilden. Alle aus der Äußerung gewonnenen SSUs werden somit miteinander verlinkt, um eine einheitliche semantische Struktur zu erzeugen.

Bei der Gewinnung der Strukturen werden zuerst die SSUs betrachtet und verlinkt, die ohne über den Umweg der hierarchischen Zuordnung direkt ermittelt werden können. Nicht immer bilden jedoch die SSUs direkt eine Verbindung aus, sondern es können auch Oberkategorien in der wurzelbildenden SSU repräsentiert werden. Da die SSU immer eine möglichst umfassende und allgemeine Beschreibung der zu ihr in Beziehung stehenden Konzepte ausbildet, kommt es regelmäßig vor, dass die in der Ableitungsstruktur an der Wurzel stehende SSU auf eine Oberkategorie verweist, die aber durch die in der Äußerung vorkommenden SSUs nicht direkt verbunden werden kann. Hier werden häufig konkrete Unterkategorien durch die entsprechenden Relationen zwischen Wort und SSU aufgeführt. In diesen Fällen müssen auch die Oberkategorien der zu verlinkenden SSUs mit in den Prozess integriert werden. Aufgrund der in der entsprechenden SSU dargestellten Hierarchiebeziehung durch den Eintrag in dem *top* Element kann diese Verbindung hergestellt werden. Daher wird in einem zweiten Schritt ebenfalls geprüft, ob die Oberkategorien der SSUs zu den wurzelgebenden SSUs passen – und somit die beiden SSUs miteinander verlinkt werden können. Dieser Prozess wird in Abschnitt 9.2 genauer beschrieben, einen ersten Eindruck sollen die folgenden Beispiele geben.

Die Äußerung „hier ist das Büro von Britta“ kann in eine vollständige semantische Struktur überführt werden. Abbildung 9.1 stellt einen vereinfachten Parseprozess dieser Ableitung dar – die Unterscheidungen zwischen obligatorischen (engl. „mandatory“) und optionalen Relationen werden hier der Übersicht halber nicht vorgenommen. Um die Ähnlichkeit mit einem Ableitungsbaum zu verdeutlichen, wie er bei klassischen grammatikalischen Verfahren Verwendung findet (vgl. [Bor91]), wurde das Parse-Ergebnis in Abbildung 9.2 in einer gebräuchlichen Anordnung der Knoten dargestellt.

Zur Gewinnung der semantischen Struktur werden zuerst die SSUs den entsprechenden Wörtern zugeordnet. Das Wort „ist“ wird auf die SSU *Existence* abgebildet. Diese stellt die Wurzel der Ableitung dar. Wurzeln oder Knoten stellen Verlinkungen zu weiteren SSUs bereit, wobei auf die Wurzel selbst keine anderen SSUs verweisen. Ebenfalls besitzt die SSU *Building_subpart* wie auch die SSU *Existence* offene Verlinkungen zu weiteren SSUs. Diese SSU *Building_subpart* wird von dem Wort „Büro“ bereitgestellt und stellt einen Knoten für einen Teil der Äußerung dar. Die SSU *Existence* besitzt unter anderem einen Link zu der SSU *Object* und einen Link zur SSU *Direction*. Die SSU *Direction* kann mit der SSU des Wortes „hier“ direkt verlinkt werden.

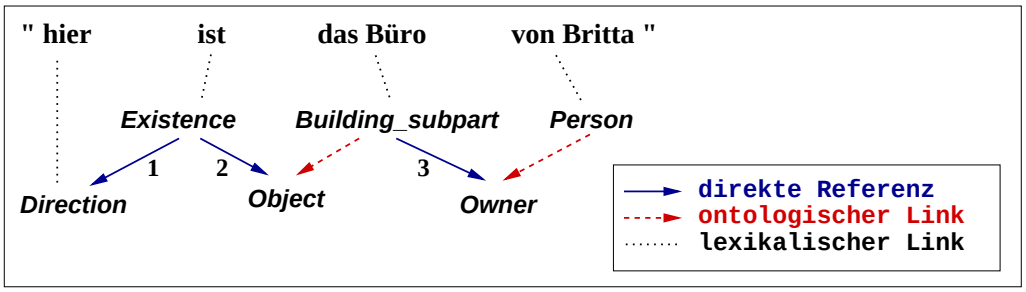


Abbildung 9.1.: Vereinfachte Darstellung des Parsevorgangs für die Äußerung „hier ist das Büro von Britta“

Die SSU *Object* kann aufgrund des Hierarchie-Wissens mit dem offenen SSU Link *Object* der SSU *Existence* verbunden werden, da ein *Building_subpart* eine Unterklasse von *Object* ist. Hierfür werden die Hierarchiebeziehungen mit einbezogen aus dem Eintrag des *top*-Elements. Das Objekt „Büro“ besitzt eine offene SSU zu einer SSU *Definite*, die den Artikel „das“ verbindet und zu einer SSU *Owner* verweist – hier wird die *Person* „Britta“ einbezogen.

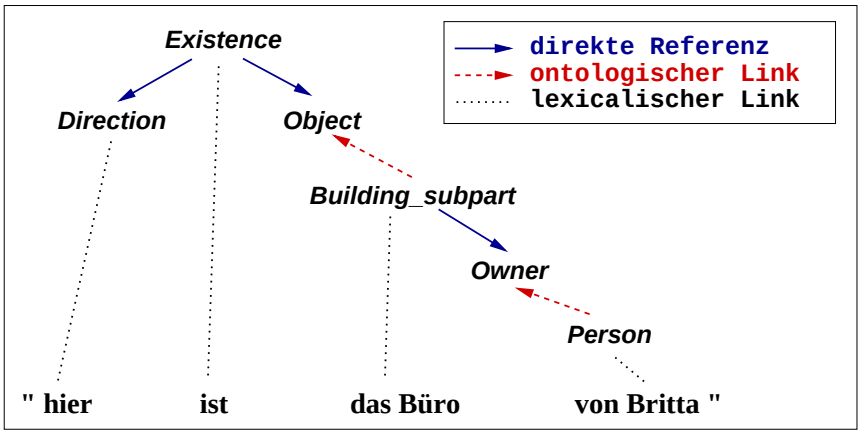


Abbildung 9.2.: Vereinfachte Darstellung der Ableitungsstruktur für die Äußerung „hier ist das Büro von Britta“

Aus dem Ergebnis des Parsevorgangs wird eine XML-Struktur erstellt, die die Ableitungsstruktur der Äußerung darstellt und dem Dialogmanager das benötigte semantische Wissen aus der Äußerung übermittelt. Die XML-Struktur für die hier beispielhaft analysierte Äußerung ist in Abbildung 9.10 dargestellt. Die Äußerung „schau – die Pflanze neben der Tür“ wird auf gleiche Weise analysiert und in eine kohärente Struktur überführt. Der Parsebaum ist in Abbildung 9.3 dargestellt.

Mit Hilfe des Lexikons und der SSUs werden aus einer Äußerung semantische Parsebäume generiert. Hierbei ist es möglich, dass für eine Äußerung nicht nur eine Ableitungsstruktur generiert werden kann, sondern eine Menge von Ableitungsstrukturen oder Teilbäumen. Nicht alle Ergebnisse ergeben daher einen Sinn oder sind als gleichwertig zu betrachten. In Kapitel 10 werden

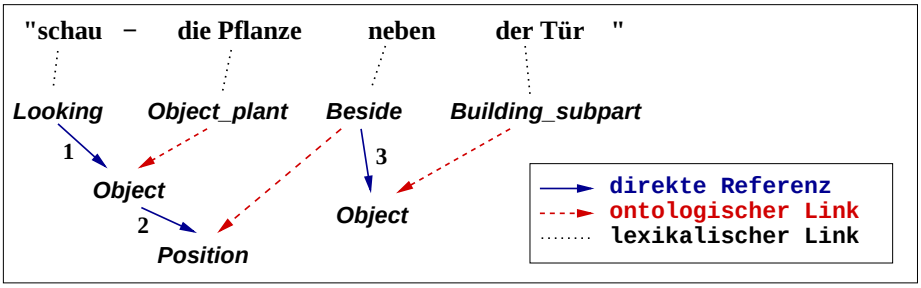


Abbildung 9.3.: Vereinfachte Darstellung des Parsebaums für die Äußerung „schau – die Pflanze neben der Tür“

die möglichen Ergebnisse ausführlicher beschrieben, eine Darstellung für mehrere Teilbäume als Ergebnis des Verarbeitungsprozesses ist z. B. in Abbildung 10.5 oder 10.13 dargestellt. Es ergibt sich einerseits die Frage, wie das beste oder die besten Ergebnisse herausgefiltert werden können, um dann die sinnvollsten Strukturen an den Dialogmanager weiterzugeben. Andererseits besteht das Problem, dass bei der Auswahl aller Ableitungsmöglichkeiten der Suchraum exponentielles Wachstum besitzt, was für die Anforderungen an ein Echtzeitsystem nicht tragfähig ist. Die zweite ebenso wichtige Frage lautet daher, wie der Suchraum soweit eingeschränkt werden kann, dass nicht alle möglichen Wege durchlaufen werden müssen, um zu den besten Ergebnissen zu kommen. Hierfür müssen Heuristiken eingesetzt werden, um die Suche soweit einzuschränken, dass die Verarbeitung in Echtzeit gewährleistet werden kann, aber dennoch die besten Ergebnisse berechnet werden können. Diese zentralen Fragestellungen werden im folgenden Abschnitt behandelt.

9.2. Der Parse-Mechanismus im Detail

Der Parse-Vorgang ist in drei Schritte unterteilt: Im ersten Schritt sucht ein *Scanner* zu jedem Wort den entsprechenden Lexikon-Eintrag und löst durch Homonyme hervorgerufene Mehrdeutigkeiten auf. Im zweiten Schritt versucht ein Verlinkungs-Prozess, möglichst alle offenen SSU-Verknüpfungen zu verlinken und so aus den Wortketten Parsebäume zu erzeugen. Im dritten und letzten Schritt werden die Parsebäume mit Hilfe einer Bewertungsfunktion evaluiert und der Beste wird als Ergebnis ausgewählt.

9.2.1. Der Scanner – Auflösung von Homonymen

Zuerst wird für jedes Wort der entsprechende Lexikon-Eintrag gesucht. Wird kein Eintrag gefunden, muss das Wort als unbekannt markiert werden. Wenn ein Eintrag im Lexikon gefunden wurde, wird die zugehörige SSU als Attribut des Wortes gespeichert. Es können aber auch im Falle eines Homonyms gleich mehrere Lexikon-Einträge gefunden werden. Um den eigentlichen Verlinkungsprozess nicht unnötig komplex gestalten zu müssen, wird die Mehrdeutigkeit von

Homonymen bereits in diesem ersten Verarbeitungsschritt aufgelöst. Dazu wird bei einem Auftreten eines Homonyms die gesamte Wortkette inklusive SSU-Attribute kopiert, wobei in den beiden resultierenden Wortketten jeweils einer der beiden möglichen Lexikon-Einträge verwendet wird. Wenn der Scanner beispielsweise auf das Wort „can“ trifft, werden zwei Wortketten erzeugt: In der einen hat das Wort „can“ die SSU *Ability* und in der anderen Wortkette die SSU *Kitchen_object*. Das Wort „can“ ist im *Home-Tour*-Szenario ein Zweifach-Homonym, d. h. es gibt zwei Lexikon-Einträge für „can“. Bei Dreifach-Homonymen müssen folglich drei Wortketten erzeugt werden, um alle Möglichkeiten zu erfassen. Noch ungünstiger wird die Situation, wenn in einer Äußerung gleich mehrere Homonyme auftreten. Dann muss für jede Kombination von Möglichkeiten jeweils eine Wortkette erzeugt werden. In Abbildung 9.4 ist der Scanner als Flussdiagramm dargestellt.

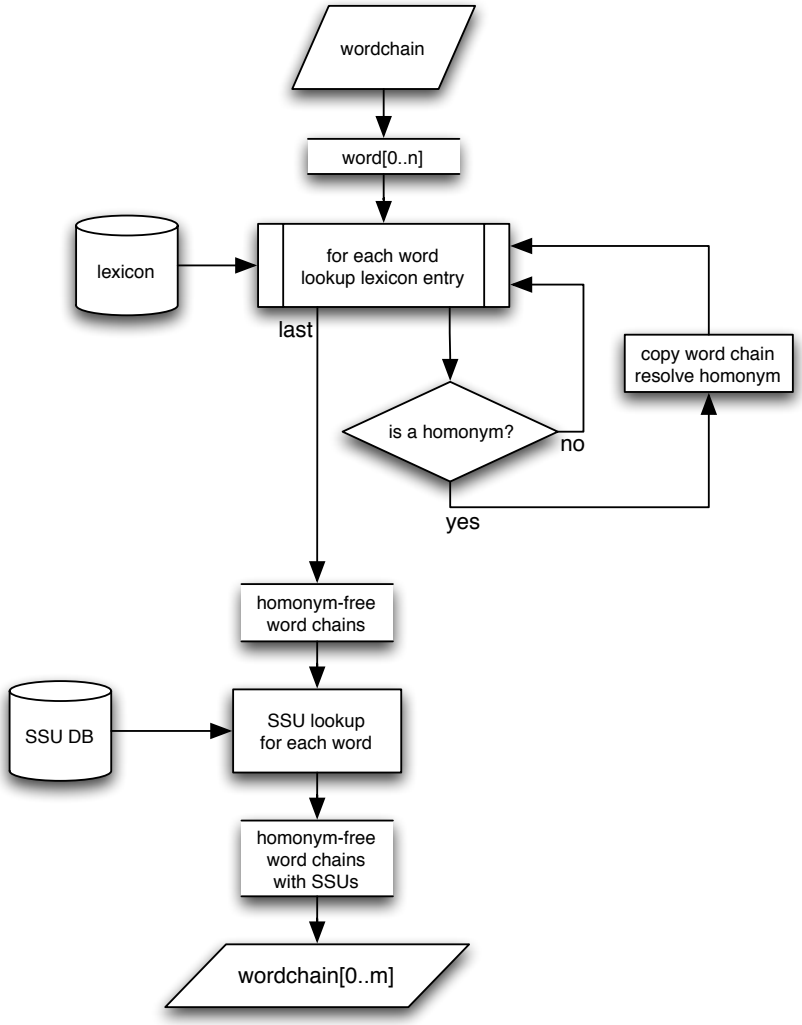


Abbildung 9.4.: Flussdiagramm des Scannerprozesses

Die Laufzeit des Scanners ist $O(2^h)$, wobei h für die Anzahl der Homonyme steht, welche nicht mit der Anzahl der Wörter verwechselt werden sollte. In der Praxis stellt sich die Situation wesentlich gutmütiger dar, denn der überwiegende Teil der Wörter in einer Äußerung sind keine Homonyme. Sie sind Sonderfälle und stellen eine Ausnahme dar, die bei gesprochener Sprache im zugrunde liegenden Szenario immerhin so häufig auftritt, dass sie bei der automatischen Interpretation keinesfalls ignoriert werden darf. Im Praxisbetrieb hat die Elimination von Mehrdeutigkeiten durch Homonyme bisher keine Laufzeitprobleme verursacht (siehe 10.3).

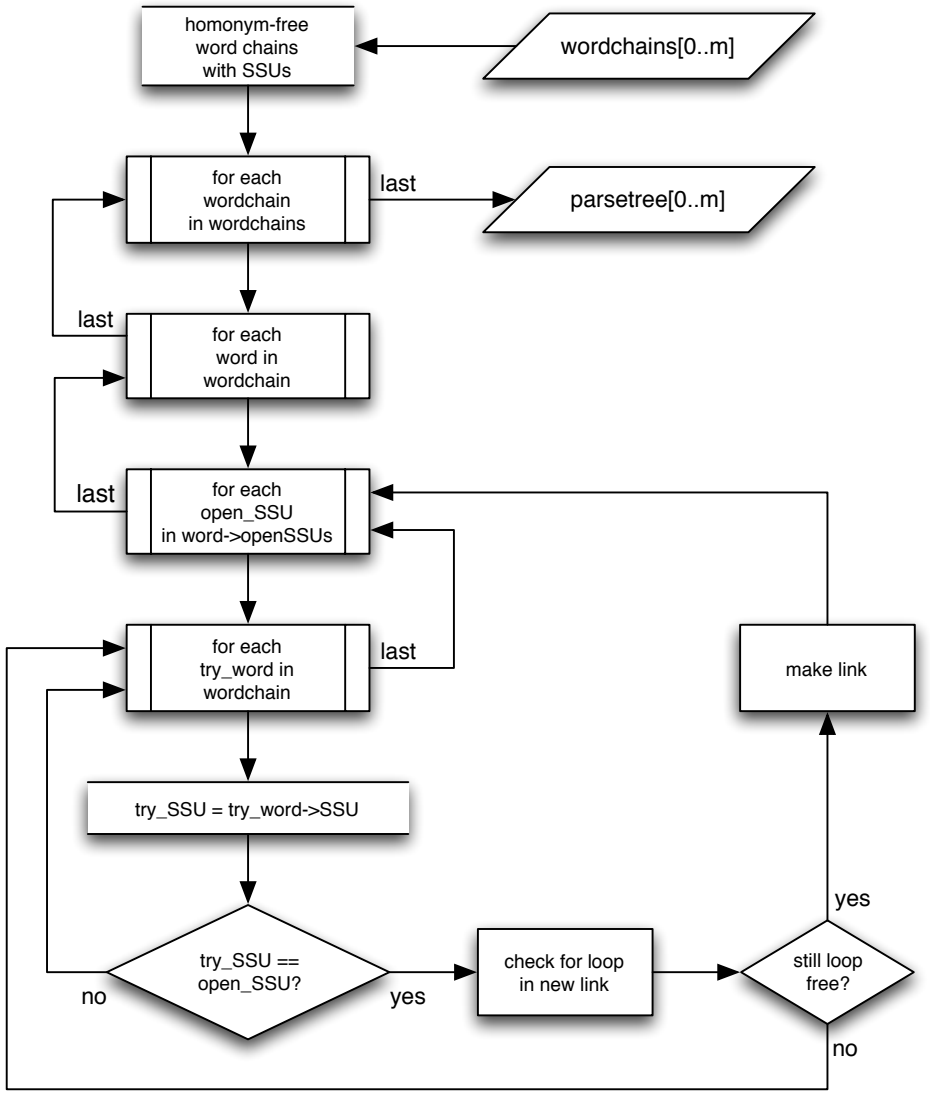
9.2.2. Der Verlinkungsprozess

Der Verlinkungsprozess hat als Eingabestrom eine Liste von Wortketten. Den einzelnen Wörtern ist jeweils bereits ein Datensatz mit der entsprechenden SSU zugeordnet. Für jedes Wort müssen nun für die bereits erwähnten offenen SSU-Verbindungen Wörter mit passenden SSUs innerhalb der Wortkette gefunden werden. Dies geschieht innerhalb mehrerer geschachtelter Schleifen, wie in Abbildung 9.5 dargestellt. Das Flussdiagramm und das Pseudocode-Fragment zeigen allerdings für eine bessere Übersichtlichkeit nur eine vereinfachte Form des Mechanismus.

Einerseits gibt es zwei verschiedene Arten von offenen SSU-Verbindungen: obligatorische und optionale. Dies spielt zwar vor allem bei der im nächsten Abschnitt beschriebenen Scoring-Funktion eine Rolle, aber auch beim Verlinkungsprozess können unterschiedliche Ergebnisse entstehen, je nachdem nach welcher Art von SSU-Verbindungen zuerst gesucht wird.

Andererseits gibt es bei den SSUs das Konzept der hierarchisch angeordneten Oberkategorien, auf das bei der Beschreibung des Parse-Vorgangs bisher noch nicht eingegangen wurde. Wenn ein Wort mit einer passenden SSU gesucht wird, kann die SSU entweder direkt mit der gesuchten übereinstimmen oder mit einer der Oberkategorien dieser SSU, welche ebenfalls durchsucht werden müssen. Es hat sich gezeigt, dass es sinnvoller ist, zunächst für alle offenen SSU-Verbindungen nach passenden Worten jeweils mittels ihrer nativen SSU – also auf der untersten Kategorie-Ebene – zu suchen, und erst danach in den Oberkategorien.

Wenn ein Wort mit einer passenden SSU gefunden wurde, muss zunächst überprüft werden, ob durch eine Verlinkung ein Zyklus oder Schleife im Graphen entstehen würde. Theoretisch kann schon eine Schleife in einer Ableitung entstehen, wenn eine SSU auf eine andere SSU verweisen kann, die wiederum auf die erstgenannte verlinkt. Dies könnte z. B. in der Äußerung „stelle den Teller auf den Tisch“ auftreten. Hier enthält die Objektbeschreibung von „Teller“ die Information einer bestimmten Position. Die Position wiederum wird bestimmt durch die Angabe eines weiteren Objektes, „Tisch“. Würde die SSU Position nicht auf „den Tisch“ verweisen, sondern auf „den Teller“, entstünde ein Zyklus im Graphen. Abbildung 9.6 zeigt die theoretisch möglichen Verlinkungen. Wenn keine Schleife entsteht, d. h. wenn die Baumstruktur durch die neue Kante erhalten bleibt, kann das Wort verlinkt werden und die Suche setzt sich mit der nächsten offenen SSU-Verbindung fort. Der Verlinkungsprozess hat für eine einzelne Wortkette eine polynomielle Laufzeit.




```
foreach wordchain in wordchains
    foreach word in wordchain
        open_SSUs = word->open_SSUs
        foreach open_SSU in open_SSUs
            foreach try_word in wordchain
                try_SSU = try_word->SSU
                if (try_SSU == open_SSU)
                    makelink(word, open_SSU, try_word)
                    last
```

Abbildung 9.5.: Flussdiagramm und Pseudocode des Verarbeitungsprozesses

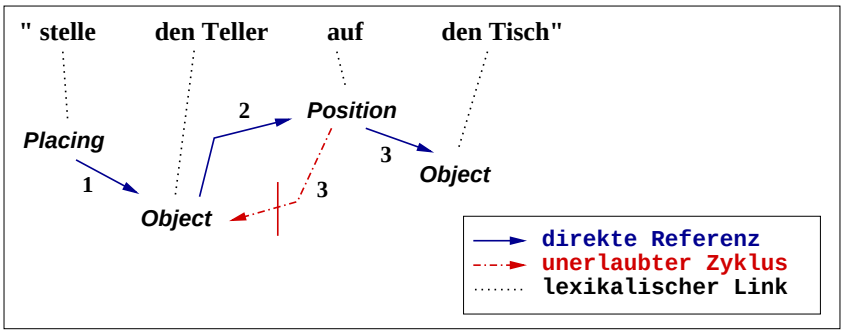


Abbildung 9.6.: Flussdiagramm des Scannerprozesses

9.2.3. Scoring – Evaluation der Parsebäume

Das Ergebnis des Verlinkungprozesses ist eine von der Anzahl der Homonyme abhängige Menge von Parsebäumen. Aus dieser Menge muss der „beste“ Parsebaum ausgewählt werden. Dies geschieht mit Hilfe einer Scoring-Funktion. Die Scoring-Funktion bewertet einen Parsebaum anhand von Kriterien wie Anzahl der Wurzeln, Anzahl der nicht-verlinkten Wörter, Anzahl der Verlinkungen, usw.

Die Scoring-Funktion dient allerdings nicht nur der Auswahl des besten Parsebaums, sondern auch einer qualitativen Beurteilung des Ergebnisses. Anhand dieser Beurteilung muß entschieden werden, wie weit dem Ergebnis der semantischen Sprachanalyse vertraut werden kann. Bei eingeschränktem oder unzureichendem Vertrauen können eventuell weitere Informationsquellen hinzugezogen werden oder das Robotersystem muss sich durch eine Rückfrage absichern.

Eine sehr einfache Bewertungsstrategie, die in der ersten Version eingesetzt wurde und bereits sehr gute Ergebnisse liefert, ist die Bewertung des Ergebnisses nach der Anzahl der verlinkten Wörter (vgl. Kap. 10.3). Das Ergebnis wird mit 100% korrekt bewertet, wenn alle Wörter im Ergebnis verlinkt werden konnten. Bei bis zu 50% verbundenen Wörtern, wird es als nicht verstanden bewertet und die Werte dazwischen als teilweise verstanden. Diese Bewertungen wurden an den Dialogmanager weitergereicht, so dass er darauf entsprechend reagieren konnte.

Alternative Bewertungsstrategien, die bisher nur im Explorations-Werkzeug für die komparative Bewertung von Parsebäumen integriert und noch nicht auf der BIRON-Plattform getestet wurden, schließen weitere Kriterien mit ein. Hier werden neben der Anzahl der verlinkten Wörter zusätzlich die obligatorischen Verlinkungen etwas stärker gewichtet als die optionalen. Ebenfalls wird besser bewertet, wenn die SSU eines Wortes direkt verlinkt werden konnte, ohne Umweg über die Oberkategorie.

Dabei hat sich herausgestellt, dass es von Vorteil ist, getrennte Scoring-Funktionen für die Auswahl des „richtigen“ Parsebaumes einerseits (komparative Evaluation) und die Bewertung der Vertauenswürdigkeit des Ergebnisses andererseits (qualitative Evaluation) zu unterhalten. Diese Vorgehensweise hat den großen Vorteil, dass sich die Kriterien für die komparative und quantitative Evaluation unterscheiden oder unterschiedlich gewichtet werden können. Dadurch ist es

möglich, Anpassungen an der komparativen Evaluation, die massiv am Parse-Ergebnis beteiligt ist, vornehmen zu können, ohne Seiteneffekte auf die quantitative Evaluation befürchten zu müssen.

9.2.4. Suchstrategien

Bei der Betrachtung des Verlinkungsprozesses sollte unbedingt beachtet werden, dass stets ein möglicher Parsebaum von vielen möglichen erzeugt wird. Es muss sich dabei nicht zwangsläufig um den gesuchten Parsebaum handeln. Welcher Parsebaum generiert wird, hängt maßgeblich davon ab, in welcher Reihenfolge nach Worten mit passenden SSUs gesucht wird. Theoretisch wäre es vorstellbar, alle möglichen Parsebäume zu erzeugen und dann den besten mittels der Scoring-Funktion auszusuchen. Der resultierende Suchraum wächst jedoch exponentiell mit der Länge der Wortkette und ist deshalb auch schon bei relativ kurzen Äußerungen so groß, dass eine Berechnung in Echtzeit sehr fragwürdig erscheint.

Daher muss der Suchraum für eine praxisgerechte Implementierung eines sprachverstehenden Systems massiv eingeschränkt werden. Dies wurde durch ein heuristisches Verfahren realisiert, das sich in ersten Versuchen bereits bewährt hat. Es wurden mehrere Suchstrategien für den Verlinkungsprozess betrachtet und miteinander verglichen. Es ist auch durchaus möglich gleichzeitig mit verschiedenen Suchstrategien im Rahmen des Verlinkungsprozesses mehrere unterschiedliche Parsebäume zu erzeugen und mittels der Scoring-Funktion zu selektieren. Da allein durch die Behandlung von Homonymen ohnehin oftmals mehrere Parsebäume generiert werden, stellt dies einen überschaubaren Mehraufwand dar.

Um bei der Implementierung des Verlinkungsprozesses möglichst schnell Ergebnisse zu erhalten, wurde zunächst die denkbar einfachste Suchstrategie verwendet: Die Suche von links nach rechts über die gesamte Wortkette. Obwohl sich leicht Wortketten konstruieren lassen, die mit dieser naiven Strategie nicht korrekt geparst werden können, hat sich in der Praxis diese Suchstrategie bereits als erfolgreich erwiesen (vgl. Kap. 10.3).

Die einfache synthetische Wortkette „the flower and the big plant“ wird vom naiven Ansatz jedoch nicht korrekt analysiert. Da für das Objekt „flower“ nach einem möglichen Attribut *Größe* (SSU *Dimension*) von links nach rechts über die gesamte Wortkette gesucht wird, verlinkt der Parser die Wörter „flower“ und „big“, was nicht der Semantik der Äußerung entspricht. Aus den Experimenten ging jedoch hervor, dass die Probanden in Äußerungen mit mehreren Objekten in der Regel gleiche Diskriminierungsmerkmale für beide Objekte verwendet haben und daher diese künstlich konstruierten Äußerungen nicht vorkamen.

Dennoch erscheint es sinnvoll, eine Suchstrategie zu entwickeln, mit der es möglich ist, auch die oben genannte synthetische Äußerung korrekt zu analysieren. Anstatt die Wortkette stur von links nach rechts nach einem Wort mit einer gesuchten SSU abzusuchen, sollte zunächst in der direkten Nachbarschaft des Ausgangswortes gesucht werden. Diese Proximal-Suche wurde zunächst innerhalb eines Explorations-Werkzeugs implementiert, welches entworfen wurde, um auch ohne die BIRON-Plattform verschiedene Suchstrategien analysieren und miteinander vergleichen zu können (siehe Kap. 9.5).

9.3. Verarbeitung von Spontansprache

Das hier beschriebene Verfahren ist besonders geeignet, um spontansprachliche Äußerungen zu verarbeiten. Spontansprachliche Phänomene bestehen vor allem aus Wortwiederholungen, der Verwendung von Füllwörtern, Abbrüchen, Neubeginn, Pausen und dem Zusammenfügen mehrerer Diskurssegmente. Aufgrund der Freiheit der Äußerung werden vermehrt Wörter verwendet, die dem System unbekannt sind. Ebenso besitzen Äußerungen zum Teil eine recht freie Wortstellung, die nicht immer den grammatikalischen Regeln der Schriftsprache folgt.

Für die Verarbeitung der Äußerungen ist die Nähebeziehung relevant, aber nicht die Reihenfolge der Satzbestandteile. Daher können Varianten eines semantischen Inhaltes durch verschiedene Äußerungen von den Interaktionspartern bereitgestellt werden, ohne dass es zu Problemen in der Verständigung kommt. Beispielsweise führen Äußerungen wie „gieße die rote Blume“, „die rote Blume gießen“ oder „die Blume äh, die rote, gießen“ zu der gleichen semantischen Interpretation (siehe Abb. 9.7). Ein weiteres Beispiel sind die folgenden Varianten: „die Tasse auf dem Tisch“, „auf dem Tisch – die Tasse“ und „die Tasse – die auf dem Tisch“. Diese Äußerungen führen ebenfalls jeweils zum gleichen Ergebnis.

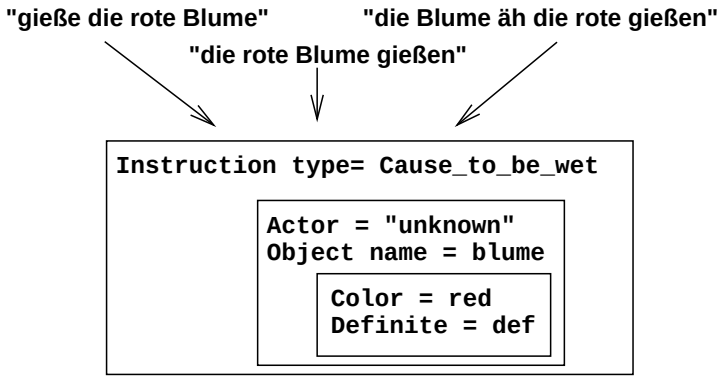


Abbildung 9.7.: Ergebnis der Äußerungsvarianten von „gieße die rote Blume“

Wortwiederholungen sind gerade im *Hometour-Korpus* ein häufiges Phänomen (vgl. Kap. 6.2). Beispiele dafür sind die Äußerungen „and look here – a cup a cup“, „this is a book is a book“ oder „look is a cup – cup“. Diese werden problemlos vom Mechanismus mit verarbeitet, bzw. der Mechanismus ignoriert doppelte Wörter. Die Interpretation der Äußerung wird hier insgesamt etwas weniger gut bewertet, da es ebenso wahrscheinlich ist, dass hier eine Fehl-Erkennung des Spracherkenners vorliegt, z. B. wenn das Ende der Äußerung falsch berechnet wurde und Umgebungsrauschen in die Erkennung einbezogen wurde oder die Äußerung nicht vollständig erkannt wurde. Füllwörter werden im Mechanismus genauso behandelt wie doppelte Wörter. Sie werden im Mechanismus weitestgehend ignoriert, falls sie nicht schon bereits im Vorfeld vom Spracherkenner herausgefiltert wurden.

Äußerungen mit Abbrüchen werden, soweit wie sie vorhanden sind, verarbeitet. Fehlen zentrale Informationen, wird das aufgrund fehlender obligatorischer Verlinkungen einer SSU sichtbar. Diese Informationen müssen dann vom Dialogsystem angefragt werden.

Ein Neubeginn mitten im Satz trat in den Studien der Mensch-Maschine-Kommunikation so gut wie gar nicht auf. Dennoch kann auch diese Art der Spontansprache auftreten. Eine mögliche Äußerung wäre z. B. „kannst du – Achtung, die Tasche fällt!“. In diesem Fall wird die Ableitung genommen, die die meisten Informationen enthält und somit die beste Bewertung bekommt. Dennoch ist die Gewinnung der korrekten semantischen Interpretation etwas problematischer. Es ist nicht immer klar, welche Informationen wirklich zusammengehören, ob die Ableitung korrekt ist oder ob ggf. eine Fehl-Erkennung vorliegt. Der Mechanismus vergibt diesen Interpretationen meist auch eine schlechtere Bewertung und veranlasst somit indirekt den Dialog, beim Interaktionspartner genauer nachzufragen. Mögliche Fehlinterpretationen werden dadurch geringer.

Die meisten Korrekturen und Ergänzungen wie „der linke äh rechte Würfel“ oder „jetzt nimm die Tasse, die blaue“ werden korrekt verarbeitet. Hierbei unterstützt, neben den semantischen Informationen der SSUs, das Prinzip der räumlichen Nähe. Dieses besagt, dass Informationen, die räumlich näher sind, bevorzugt verlinkt werden.

In vielen Fällen erkennt der Spracherkenner das Ende einer Äußerung korrekt, auch wenn mittendrin Pausen auftauchen. Diese einzelnen Teile werden vom Verarbeitungsmechanismus so lange in der Historie gesammelt, bis die Äußerung vollständig ist. Da in der Interaktion zwischen Roboter und Mensch davon ausgegangen wird, dass es sich um einen tatsächlichen Dialog handelt, der Mensch also auf die Reaktion des Roboters wartet und umgekehrt, besteht nur eine geringe Gefahr, dass der Roboter zu viele Informationen gleichzeitig erhält, die ihn dann überfordern. Es wird eher erwartet, dass der Mensch auf die Reaktion des Roboters wartet und dann erst fortfährt. Überschreitet eine Sprechpause eine bestimmte Zeitdauer, so wird sie vom Spracherkenner als Abbruch einer Äußerung interpretiert. Er zerschneidet die Äußerung an dieser Stelle und es wird nur ein Teil der Äußerung an das Sprachverstehen weitergeleitet. Nur dieser Teil wird dementsprechend durch den Mechanismus verarbeitet und auf gleiche Art und Weise behandelt wie Satzabbrüche.

Bei der Äußerung mehrerer Diskurssegmente wie „hallo - schau mal her“ wird genauso vorgegangen wie bei einem Neubeginn. Hier wird die Ableitung mit den meisten Informationen ausgewählt, in diesem Fall der Teil „schau mal her“. Können zu viele Wörter nicht zusammengefügt werden, erhält die Interpretation wiederum eine schlechtere Bewertung und der Dialogmanager startet eine Rückfrage. Der Grundgedanke ist, dass lieber einmal mehr gefragt wird als dass der Roboter eine falsch erkannte Anweisung ausführt. Die Analyse der erkannten Äußerungen ist in Kapitel 10 ausführlicher beschrieben.

9.4. Anaphernresolution

Um eine Anaphernresolution zu realisieren, werden Informationen aus unterschiedlichen Quellen benötigt. Das Wissen aus dem Sprachverstehen alleine reicht dafür nicht aus. Zunächst muss für die aktuelle Äußerung festgestellt werden, ob überhaupt eine Anapher vorliegt. Zum Beispiel könnte das Demonstrativpronomen „diese“ auf eine Anapher oder aber auch auf eine Zeigegeste

hindeuten. Leider ist ohne weiteres Wissen nicht zu erkennen, ob es eine Referenz auf die Szene darstellt oder ein Verweis auf bereits durch Sprache bekanntes Wissen. Nur wenn keine Geste verwendet wurde, trägt das entsprechende Wort eine anaphorische Bedeutung. Um die eventuell vorhandene Anapher erkennen und ggf. auflösen zu können, sind daher weitere Informationen aus der Szene und aus der Historie des Dialoges relevant.

Dennoch kann das Sprachverstehen wesentlich zur Auflösung der Anaphern beitragen. Es gibt Hinweise auf eine mögliche anaphorische Verwendung eines Wortes und reicht sie an den Dialogmanager weiter.[1] Wörter, die auf eine Anapher verweisen, sind u. a. Pronomen, insbesondere Demonstrativpronomen. Letztere werden im System als Homonyme betrachtet – sie werden sowohl als Link auf die SSU *Maybe_gesture* als auch auf die SSU *Anaphoric* gespeichert. Die Pronomen an sich werden nur als *Anaphoric* interpretiert, sie werden in der Regel nicht für die Referenz auf Szeneinformationen genutzt. Der Verarbeitungsmechanismus kann die SSU dann nutzen, um entsprechend Hinweise auf mögliche Anaphern oder Gesten zu geben.

Zur Zeit findet im Robotersystem BIRON noch keine Anaphernauflösung statt, sie ist jedoch zukünftig geplant. Der Zugriff auf visuelle Informationen und auch auf die Historie der Sprache geschieht im Robotersystem über die Dialogsteuerung, hier ist die Entscheidungsfindung des Analyseverfahrens angesiedelt. Wenn ein Hinweis auf eine Anapher kommt, kann der Dialogmanager zunächst eine Anfrage an das Szenemodell oder an die Aufmerksamkeitssteuerung für Objekte schicken, um zu klären, ob visuelle Informationen zur Auflösung dienen können. Wird von dort keine Information geliefert, findet die Anaphernresolution aufgrund der sprachlichen Historie statt. Der Dialogmanager entscheidet letztendlich, welche Informationsquellen genutzt werden können, da nur er Zugang zu den Quellen und das genaue Wissen über die Dialogsituation besitzt. Die Auflösung selbst kann wiederum mit Hilfe des Verarbeitungsmechanismus stattfinden. Im Ausblick (Kap. 11.2) werden verschiedene Verfahren zur Anaphernresolution und Ellipsen-Auflösung beschrieben.

9.5. Das Explorations-Werkzeug – Debugger für Suchstrategien, Lexikon und SSU-Datenbank

Die Sprachverstehenskomponente innerhalb der Biron-Plattform wurde unter besonderer Berücksichtigung der unkomplizierten Erweiterbarkeit von Lexikon und SSU-Datenbank entworfen. Ebenso sind die Suchstrategien für den Verlinkungsprozess oder die Kriterien der Scoringfunktion relativ einfach austauschbar. Dennoch stellt es einen recht großen Aufwand dar, kleine Anpassungen oder Verbesserungen des Lexikons oder der SSU-Datenbank direkt auf der Biron-Plattform zu testen, weil in diesem Fall immer nur der Gesamtprozess der Mensch-Roboter-Interaktion getestet werden kann. Daher müssen Testszenarien entwickelt werden, mit denen

[1] Zur Zeit wird nur genau eine Lösung an den Dialogmanager gesendet (mit Präferenz der Gesten vor den Anaphern), es ist jedoch problemlos möglich, mehrere Ergebnisse zu senden, die der Dialogmanager parallel nutzen kann.

festgestellt werden kann, ob die Änderungen oder Verbesserungen zu dem gewünschten Ergebnis führen. Aber gerade in Situationen, in denen es mehrere Möglichkeiten gibt, einen Lexikon- oder SSU-Datenbank-Eintrag zu gestalten, wäre es wünschenswert, ohne aufwendige Tests und damit schneller und einfacher herausfinden zu können, wie sich die vorgenommenen Veränderungen auswirken.

Vor diesem Hintergrund wurde das Explorations-Werkzeug entwickelt. Das in Perl implementierte kommandozeilenorientierte Utility nimmt Äußerungen per Tastatureingabe entgegen und gibt die generierte Interpretation in Form eines oder mehrerer Parsebäume aus. Genau wie die Implementierung innerhalb der Biron-Plattform liest auch das Explorations-Werkzeug sowohl Lexikon als auch SSU-Datenbank jeweils in der gleichen XML-Notierung ein.

Das Explorations-Werkzeug unterstützt folgende Befehle:

`\lex <word-name>`	zeigt den Lexikon-Eintrag zu `<word-name>`
`\ssu <SSU-name>`	zeigt die gesamte SSU `<SSU-name>`
`\parsefuncs`	listet die verfügbaren Such-Strategien auf
`\use <function-name>`	wählt eine Such-Strategie aus

Alle anderen Eingaben werden als Äußerung aufgefasst und interpretiert. Die Bereitstellung der verschiedenen Suchstrategien soll ermöglichen, dass unterschiedliche Varianten des Verarbeitungsmechanismus ausgetestet werden können. Für die heuristische Suche scheint keine allgemeine Universallösung zu existieren, die für jede denkbare Äußerung die richtige Interpretation liefert. Daher kann innerhalb des Explorations-Werkzeug getestet werden, wann welche Methode besonders gut geeignet ist und wann fehlerhafte Interpretationen generiert werden. Je nachdem, wieviel Zeit für den Verarbeitungsmechanismus zur Verfügung steht, können im System auch mehrere Mechanismen parallel laufen, um so aus den verschiedenen Ableitungen das beste Parse-Ergebnis herauszufiltern. Ebenfalls können lokale Such-Präferenzen der einzelnen SSUs selbst berücksichtigt werden (siehe Kap. 11.2). In der hier beschriebenen Version des Explorations-Werkzeugs wurden folgende Suchstrategien implementiert:

lr_lr_rel_maybe liest die Wortkette von links nach rechts und durchsucht die Wortkette von links nach rechts nach SSUs. Es werden für ein Wort erst nach obligatorischen (rel) und danach nach optionalen (maybe) SSUs gesucht. Diese Suchstrategie repräsentiert den naiven Ansatz und sie wurde in der ersten Implementierung auf der Biron-Plattform verwendet und ist angelehnt an die zeitliche Reihenfolge sprachlicher Äußerungen.

lr_prox_rel_maybe liest die Wortkette von links nach rechts und durchsucht die Wortkette nach SSUs mit der Proximalsuche, d. h. erst wird die direkte Nachbarschaft ausgehend vom Wort mit den gesuchten SSUs und dann in größer werdender Distanz. Auch hier wird zuerst nach obligatorischen und danach nach optionalen SSUs gesucht. Dieser Ansatz liefert mit synthetischen Äußerungen deutlich bessere Ergebnisse als der naive Ansatz. Er verfolgt die Idee, dass sich in sprachlichen Äußerungen die semantische Zusammengehörigkeit durch eine räumliche Nähe der Wörter innerhalb der Wortkette zeigt.

lr_lr_maybe_rel funktioniert wie `lr_lr_rel_maybe` mit dem Unterschied, dass erst nach optionalen und dann nach obligatorischen SSUs gesucht wird, wodurch andere Interpretationen generiert werden können. Diese Suchstrategie eignet sich besonders, um beim Entwurf neuer SSUs die Ausprägung einer offenen SSU-Verbindung als obligatorisch oder optional zu untersuchen.

rl_rl_rel_maybe liest die Wortkette von rechts nach links und durchsucht die Wortkette von rechts nach links nach SSUs. Attribute von Objekten stehen in der Regel vor diesen, wodurch die Suche von rechts nach links motiviert wird. Beispielsweise befindet sich in der Äußerung „die blaue Blume“ das Attribut *Farbe* links vom Objekt.

lr_rl_rel_maybe liest die Wortkette von links nach rechts und durchsucht die Wortkette von rechts nach links nach SSUs. Ansonsten wie `lr_lr_rel_maybe`. Diese Suchstrategie stellt eine Kombination von `lr_lr_rel_maybe` und `rl_rl_rel_maybe` dar, um einerseits der zeitlichen Reihenfolge sprachlicher Äußerungen und andererseits den in der Regel vor Objekten auftretenden Attributen Rechnung zu tragen.

In Abbildung 9.8 ist ein Auszug aus einer typischen Sitzung mit dem Explorations-Werkzeug dargestellt. Das Explorations-Werkzeug hat sich als gutes Hilfsmittel bewährt, um kleine Anpassungen an Lexikon oder SSU-Datenbank zu testen oder um verschiedene Suchstrategien miteinander zu vergleichen.

9.6. Integration in die Roboterplattform BIRON

Der Prozess der Sprachverarbeitung spiegelt sich durch eine Vielzahl von Modulen wider, die miteinander agieren und aufeinander abgestimmt sein müssen, um ein Robotersystem mit tiefergehenden kommunikativen Fähigkeiten hervorzubringen. In diesem Abschnitt wird erläutert, wie das sprachverstehende System mit den in Verbindung stehenden Komponenten interagiert, wie es sich in den Gesamtaufbau der Dialogarchitektur des Roboters BIRON einfügt und wie die Informationen der Sprachdaten im System fließen.

Das sprachverstehende System steht direkt in Verbindung zur Spracherkennung, von der es die zu analysierende erkannte Äußerung erhält, und zum Dialogsystem, das die Ergebnisse der Sprachverstehenskomponente für die Interaktion mit dem Kommunikationspartner des Roboters nutzt. Der Themendetektor nutzt ebenfalls die Ergebnisse aus dem Sprachverstehen, wobei er vor allem über den Dialogmanager kommuniziert und mit der Sprachverstehenskomponente nur indirekt in Verbindung steht.

9.6.1. Konvertierung der Ergebnisse aus dem Spracherkenner

Für das Robotersystem wird die in [Haa04] beschriebene sprecherunabhängige HMM-basierte Spracherkennung verwendet (vgl. Kap. 5.6.1). Zur Zeit wird dafür noch eine Grammatik verwen-

```
Reading lexicon (./lexicon.xml)... (1393) OK
Reading frameDB (./frames.xml)... (152) OK
>\parsefuncs
lr_lr_rel_maybe   : Scan and search left to right, rel SSUs first, then maybe SSUs
lr_prox_rel_maybe : Scan left to right, proximity search, rel SSUs first, then maybe SSUs
lr_lr_maybe_rel   : Scan and search left to right, maybe SSUs first, then rel SSUs
rl_rl_rel_maybe   : Scan and search right to left, rel SSUs first, then maybe SSUs
lr_rl_rel_maybe   : Scan left to right, search right to left, rel SSUs first, then maybe SSUs
>\use lr_lr_rel_maybe
>the flower and the red cup
combinations found by de-homonymizer: 2
parse_trees produced by unify():      2
parsing strategy:                     lr_lr_rel_maybe
                          |
                    +------------+
                    | and:Adding |
                    +------------+
               _____________|_________________
               |                              |
    +----------------------+     +--------------------+
    | flower:Object_plant |     | cup:Object_kitchen |
    +----------------------+     +--------------------+
      __________|_______                    |
      |               |                    |
+-----------+ +--------------+    +--------------+
| red:Color | | the:Definite |    | the:Definite |
+-----------+ +--------------+    +--------------+

>\use lr_prox_rel_maybe
>the flower and the red cup
combinations found by de-homonymizer: 2
parse_trees produced by unify():      2
parsing strategy:                     lr_prox_rel_maybe
                          |
                    +------------+
                    | and:Adding |
                    +------------+
          ___________________|_____________
          |                               |
+----------------------+     +--------------------+
| flower:Object_plant |     | cup:Object_kitchen |
+----------------------+     +--------------------+
          |                    __________|_______
          |                    |                |
   +--------------+     +-----------+ +--------------+
   | the:Definite |     | red:Color | | the:Definite |
   +--------------+     +-----------+ +--------------+

>\frame Object_kitchen
$VAR1 = { 'maybe_SSUs' => [
                            'Direction'
                          ],
          'rel_SSUs' => [
                          'Color',
                          'Position',
                          'Dimension',
                          'Owner',
                          'Maybe_gesture',
                          'Definite'
                        ],
          'top_SSU' => 'Object'
        };
```

Abbildung 9.8.: Auszug aus einer Sitzung mit dem Explorations-Werkzeug

det, später soll jedoch ein unabhängiges Sprachmodell erstellt werden. Daher werden aktuell die Ergebnisse der Spracherkennung als hierarchische Baumstruktur *„part-of-speech“* (POS) ausgegeben. Diese müssen für den Sprachverstehensprozess analysiert und in eine Wortfolge transformiert werden. Besonders zu beachten ist hierbei auch, dass der Spracherkenner auf das Kommunikationsframework DACS zurückgreift, die restlichen Komponenten des Robotersystems jedoch auf das Kommunikationsframework XCF (vgl. Kap. 5.3).

Einerseits sollte die Möglichkeit bereit gestellt werden, dass das sprachverstehende System unterschiedliche Spracherkenner für seinen Prozess nutzen kann. Es soll unabhängig von der genauen Ausgabestruktur des Spracherkenners sein. Andererseits soll die Möglichkeit bestehen, den Datentransfer der Roboterkomponenten alleinig auf XCF umstellen zu können, welches, im Gegensatz zu DACS, ein geeignetes Kommunikationframework für XML-Schemata darstellt. Daher wurde die sprachverstehende Komponente nicht direkt an die Spracherkennung angebunden, sondern eine Schnittstelle dazwischen geschaltet.

Diese filtert die einzelnen Wörter der Äußerung aus der POS-Datenstruktur, die in der Programmiersprache C vorliegt, und wandelt sie in ein allgemeines XML-Schema um. Werden andere oder weitere Spracherkenner eingesetzt, so muss einzig die Schnittstelle angepasst werden, die weiteren Module können so belassen werden. Für das offene Modulkonzept der Roboterplattform stellt diese Methode eine wichtige Voraussetzung dar.

9.6.2. Der Sprachverstehensprozess im Robotersystem BIRON

Startet das Robotersystem, so werden zunächst einige Grundinformationen eingelesen: die Wahl des Verarbeitungsmechanismus[2] sowie Name und Quelle des Lexikons und der semantischen Konzeptdatenbank. Das Lexikon wird zunächst in eine Hashtabelle eingetragen, wobei Homonyme mehrere alternative Relationen zu den entsprechenden SSUs enthalten. Die Konzeptdatenbank wird ebenfalls in eine Hashtabelle eingetragen, in der die offenen Links eingetragen sind. Die Hashtabelle enthält genauer gesagt eine Datenstruktur einer SSU, die die offenen obligatorischen und optionalen Links enthält und zusätzlich einen Eintrag für die Oberkategorie. Durch die Wahl des Datenzugriffs in Form einer Hashtabelle kann sichergestellt werden, dass auch bei einem größerem Wortschatz ein schneller Zugriff erfolgen kann. Durch die Speicherung der Daten in einer externen Datenbank ist ein problemloser Austausch der Sprachdaten gewährleistet und der Wortschatz kann an die jeweiligen Aufgabenstellungen und die Umgebung des Roboters angepasst werden.

Jetzt ist die Sprachverstehenskomponente initialisiert und bereit, Eingaben zu verarbeiten und wartet auf Wortketten aus der Schnittstelle zum Spracherkenner. Sobald Informationen eingehen, startet sie den semantischen Interpretationsvorgang der Äußerungen. Für jedes Wort wird eine entsprechende Instanz oder ein Datensatz erzeugt. Sie enthält die wichtigen Informationen, um sie mit anderen Instanzen verbinden zu können. Das sind u. a. der Lexikonname, der Name der SSU, die hierarchische Zuordnung sowie die in der SSU enthaltenen obligatorischen und op-

[2] Die Vorbelegung stellt die Variante von links nach rechts dar.

tionalen Relationen zu anderen SSUs. Diese Liste der Relationen stellt die Verbindungsstelle zu den in Beziehung stehenden SSUs dar. Weiterhin hält die Instanz Behälter für die Informationen bereit, die zur Berechnung und Bewertung wichtig sind, z. B. auf wieviele Blätter oder Knoten sie verweist und ob sie eine Wurzel, einen Knoten oder ein Blatt darstellt sowie die Bewertung der Instanz. Ebenfalls ist die Stelle in der Wortkette vermerkt, um direkt auf die Nachbarn zugreifen zu können und eine lineare Berechnung der verschiedenen Varianten zu ermöglichen. Mit diesen Hinweisen können indirekte Schleifen in der Ableitungsstruktur verhindert werden. Somit kann garantiert werden, dass tatsächlich ein Parsebaum entsteht. Denn es darf nicht vorkommen, dass eine SSU auf eine SSU verweist, die diese SSU bereits enthält. Bei Homonymen wird der Vorgang mit jeweils einem Homonym entsprechend wiederholt.

Je nach Analysevariante werden nun die SSUs (oder Instanzen) in der entsprechenden Reihenfolge miteinander verbunden, bis alle Wörter durchlaufen sind und die SSUs mit ihren möglichen Relationen zu anderen SSUs verbunden sind. Letztendlich wird der Parsebaum (gefüllte Instanz) herausgesucht, der die beste Bewertung erhält. Mitunter kann es vorkommen, dass zwei Ergebnisse gleich bewertet werden, z. B. wenn Homonyme im Prozess beteiligt sind. In diesem Fall wird das Ergebnis nach den enthaltenen SSUs ausgewählt, d. h. bestimmte SSUs werden anderen SSUs vorgezogen. Bei der Wahl zwischen Anapher und Hinweis auf eine Geste wird beispielsweise ein Ergebnis mit der SSU *Maybe_gesture* vor der SSU *Anaphoric* bevorzugt, da zur Zeit Anaphern nicht aufgelöst werden. Ebenso haben Hinweise auf bestimmte Fragekonstellationen (z. B. nach Thema) Vorrang. Denkbar ist jedoch auch, beide Ergebnisse weiterzuleiten und die Entscheidung dem Dialogmanager zu überlassen, die er je nach Kenntnis der Situation gezielt treffen kann.

9.6.3. Repräsentation der Äußerungen für den Dialogmanager

Die Ergebnisse des Sprachverstehens werden in ein XML-Schema konvertiert und mittels des Kommunikationsframeworks XCF (siehe Kap. 5.3) an den Dialogmanager übertragen. Gleichzeitig erhält auch der Themendetektor diese Daten, um aus der Äußerung und dem Dialogakt die aktuelle Situation zu überprüfen und ggf. ein neues Thema (*Topic*) anzukündigen.[3]

Die Ergebnisse aus dem Sprachverstehen werden in zwei Bereiche unterteilt. Zum einen werden sogenannte Meta-Informationen in der XML-Struktur gespeichert. Dabei wird zunächst ein Zeitstempel, der den zeitlichen Beginn der Äußerung markiert, im Element *timeStamp* gespeichert. Dieser wird neben der reinen Wortfolge ebenfalls vom Spracherkenner an das sprachverstehende System weitergereicht. Desweiteren wird auch die Bewertung der semantischen Kohärenz – und somit auch die Güte der Spracherkennung – weitergeleitet (in dem XML-Element *processingStatus*). Hiernach kann der Dialogmanager entscheiden, ob – und wenn ja, welche – Rückfragen er an den Interaktionspartner stellt. Das System klassifiziert drei Kategorien: *fullunderstanding*, *partialunderstanding* und *nonunderstanding*. Für die Bestimmung der Güte wird der Bewertungsmechanismus verwendet, der die Anzahl der verlinkten Wörter mit der Anzahl der Wörter

[3] z. B. um die Spracherkennung auf ein neues Thema zu eichen

in der Äußerung insgesamt vergleicht, siehe Kapitel 9.2.3. Die genaue Zuordnung kann je nach Kontext variabel stattfinden. Es hat sich in den Experimenten jedoch herausgestellt, dass es sinnvoll ist, Äußerungen, die entweder vollständig oder zu mindestens 75% verlinkt werden konnten, als *fullunderstanding* zu bewerten. Diese Äußerungen sind semantisch kohärent und ergeben einen Sinn, der ohne weiteres zu verstehen ist. Dagegen werden Äußerungen mit mindestens 50% verlinkten Wörtern als *partialunderstanding* und Äußerungen mit weniger als *nonunderstanding* bewertet. Bei nur zum Teil verständlichen Äußerungen müssen vom System spezielle Rückfragen, unter Einbeziehung des bereits erlangten Wissens aus dem Sprachverstehen, gestellt werden. Gilt eine Äußerung als *nonunderstanding*, so konnte keine Information daraus gewonnen werden. Hierbei handelt es sich mit großer Wahrscheinlichkeit um eine vollständige Fehl-Erkennung (z. B. aufgrund von Störgeräuschen oder unbekannten Inhalten). In dem Element *asrBehavior* wird vermerkt, ob ein ungewöhnliches Verhalten vom Spracherkenner auftritt, z. B. die Daten nicht korrekt geliefert werden.

```
<utterance>
   <metaInfo>
   <timeStamp>1125573609754</timeStamp>
   <processingStatus>full</processingStatus>
   <asrBehavior>normal</asrBehavior>
   </metaInfo>
   <SemanticInfo>
      <plainText>what can you do</plainText>
      <category>query</category>
      <content>
         <unit = "Question_action">
            <name>what</name>
            <unit = "Action" >
               <name>do</name>
               <unit = "Ability" >
                  <name>can</name>
               </unit>
               <unit = "Proxy" >
                  <name>you</name>
               </unit>
            </unit>
         </unit>
      <content>
   </SemanticInfo>
</utterance>
```

Abbildung 9.9.: Auszug aus der XML-Repräsentation der Äußerung „what can you do“ – das Ergebnis des Sprachverstehensprozesses

Zusätzlich zu den übergeordneten Informationen wird die reine Äußerung als Wortfolge weitergeleitet. Wesentlich wichtiger ist jedoch die Übertragung der semantischen Interpretation der Äußerung. Besonders relevant hierbei ist die korrekte Angabe des Dialogaktes, welcher in dem XML-Element *category* gespeichert wird. Hierfür steht in einer externen Tabelle die Zuordnung

der SSUs zu den entsprechenden Dialogakten (vgl. auch Kap. 6.2). Für die Interpretation der Äußerung bestimmt die Wurzel-SSU den jeweiligen Dialogakt. In den meisten Fällen findet eine Zuordnung über die Oberkategorie der SSU statt, z. B. bei Unterklassen von *Object* oder *Action*, in einigen Fällen auch direkt, z. B. bei *Negation* oder *Confirmation*. Kann eine Wurzel-SSU keinem Dialogakt zugeordnet werden, so wird der Dialogakt mit dem Eintrag *fragment* als unbekannt markiert. Für den Dialogmanager werden diese Äußerungen als nicht-verständlich klassifiziert, und eine Rückfrage wird gestartet. Der Ableitungsbaum der Interpretation wird ebenfalls dargestellt. Die zentralen Informationen aus den SSUs, die die Wurzel darstellen, stehen dabei weiter außen. Dadurch ist gewährleistet, dass der Dialogmanager nur so tief in die Struktur hineinschauen muss, wie er die Informationen konkret für den Dialog benötigt, z. B. um in der Szene das gesuchte Objekt eindeutig zuordnen zu können.

Die Abbildung 9.9 zeigt ein Beispiel des Ergebnisses aus dem Konvertierungstool des Sprachverstehens für die Äußerung „what can you do“. Abbildung 9.10 stellt den Auszug der semantischen Informationen der Äußerung „hier ist das Büro von Britta“ dar, der ebenfalls als Parsebaum in Abbildung 9.1 dargestellt ist.

```
<SemanticInfo>
   <plainText>hier ist das Büro von Britta</plainText>
   <category>description</category>
   <content>
      <unit = Existence>
         <name>ist</name>
         <unit = Direction>
            <name>hier</name>
         </unit>
         <unit = Building_subpart>
            <name>Büro</name>
            <unit = Definite>
               <name>das</name>
            </unit>
            <unit = Owner>
               <name>Britta</name>
            </unit>
         </unit>
      </unit>
   </content>
</SemanticInfo>
```

Abbildung 9.10.: Auszug aus der XML-Repräsentation der Äußerung „hier ist das Büro von Britta“ – das Ergebnis des Sprachverstehensprozess

9.6.4. Das Dialogsystem im Überblick

An dem gesamten Dialogprozess sind mehrere Komponenten des Robotersystems BIRON beteiligt. Abbildung 9.11 zeigt eine Übersicht über die Dialogarchitektur des Robotersystems BIRON (vgl. auch Kap. 5.9).

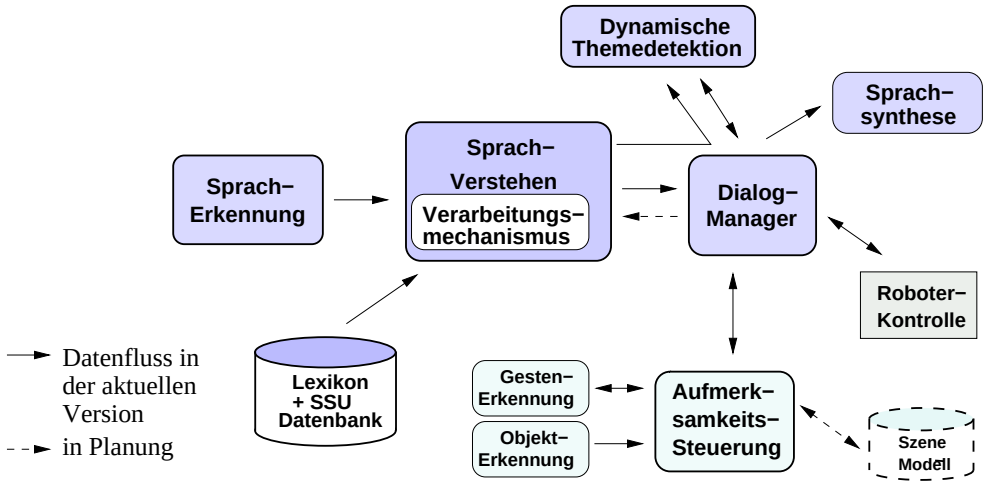


Abbildung 9.11.: Übersicht der Dialogarchitektur des Robotersystems BIRON

Wie in Kapitel 5.6.1 beschrieben erhält die Spracherkennung akustische Sprachsignale der Äußerungen, die sie in eine symbolische Repräsentationsform konvertiert. Diese symbolische Struktur wiederum dient dem Sprachverstehensprozess als Grundlage für die Analyse. Mit Hilfe der Schnittstelle zwischen Spracherkennung und Sprachverstehen wird aus den Zeichen, die als komplexe Datenstruktur vorliegen, eine Wortkette generiert, die das Sprachverstehen für die Analyse verwendet.

Die aus den Äußerungen gewonnenen semantischen Ableitungsstrukturen werden schließlich in ein XML-Schema umgewandelt und mit Hilfe des XCF-Kommunikations-Frameworks an den Dialogmanager weitergeleitet. Die Interpretationen nutzt der Dialogmanager, um Handlungsaufforderungen an die Robotersteuerung weiterzuleiten, um Szeneinformationen einzufordern oder um Antworten an den Interaktionspartner zu generieren. Um den Arbeitsaufwand so gering wie möglich zu halten, nutzt er zunächst die Dialogakte als erste Informationsquelle und nach Bedarf auch den semantischen Inhalt der XML-Struktur. Die Meta-Informationen, die ebenfalls mit der semantischen Ableitungsstruktur geschickt werden, werden ebenfalls ausgelesen und bei Bedarf verwendet (z. B. der Zeitstempel zum zeitlichen Abgleich mit einer Geste oder Bewertung der Interpretationsgüte um ggf. Rückfragen zu stellen oder situationsangemessen zu antworten). Der Dialogmanager ist quasi die Kontrollinstanz über den Dialog und die Schnittstelle zu anderen Modulen im Robotersystem (Themendetektion, Roboterkontrolle). Er leitet die entsprechenden Informationen an die anderen Module weiter und ist hauptverantwortlich für den Dialogverlauf. Die akustische Ausgabe der Antworten und Rückfragen an den Interaktionspartner erfolgt über ein Sprachsynthesetool.

Der Themendetektor analysiert die Kommunikation nach aktuellen Themen und kann so die gesamte Interaktion unterstützen. Beispielsweise kann er durch die Angabe des aktuellen Themas den Suchraum der Objekterkennung einschränken (z. B. ist das Thema „Tee-Kochen“ oder „Schreibtisch“). Der Themendetektor nutzt dafür neben den visuellen Informationen der aktuellen Situation ebenfalls die Wortkette, die Dialogakte und die semantischen Interpretationen aus dem Sprachverstehen.

9.7. Zusammenfassung

Das in diesem Kapitel beschriebene Verfahren setzt ein Konzept um, das die situierten semantischen Einheiten nutzen kann, um aus einer spontansprachlichen Äußerung eine semantische Interpretation zu gewinnen. Mit ihrer Hilfe kann eine situationsangepasste Kommunikation zwischen Roboter und Mensch stattfinden.

Darüber hinaus generiert das System ein breites Informationsspektrum aus der Äußerung: Es stellt neben der reinen semantischen Interpretation der Äußerung auch Dialogakte bereit und gibt so viele zusätzliche Informationen wie möglich, die dem Dialog hilfreich sein können. Es gibt zusätzlich Hinweise auf visuelle Informationen, auf fehlende Informationen und auf wahrscheinlich anaphorischen Sprachgebrauch (Hinweise auf Informationen aus der Sprachhistorie).

Das Verfahren berücksichtigt mehrere Rahmenbedingungen (vgl. Kap. 7), die in der Mensch-Roboter-Kommunikation zentrale Bestandteile darstellen: Es ist geeignet zur Interpretation situierter Spontansprache und kann in Robotersystemen eingesetzt werden, die in Echtzeit agieren. Zusätzlich bietet es Informationen zur Güte der Spracherkennungsleistung an.

Der Verarbeitungsprozess ist in dem Robotersystem BIRON integriert und unterstützt den Gedanken des offenen Architekturkonzeptes. Die direkten Schnittstellen sind jeweils im XML-Repräsentationsformat umgesetzt und bieten daher eine sowohl von der Programmiersprache als auch von dem konkreten Betriebssystem weitestgehend unabhängige Übertragungsmöglichkeit. Die Wissensdatenbanken werden extern verwaltet und können ausgetauscht werden, ohne dass der Verarbeitungsmechanismus in irgendeiner Weise verändert werden muss.

10. Evaluation

Für die Evaluation eines sprachverstehenden Systems existieren keine Standards, auf die zurückgegriffen werden kann. Die Anforderungen, unter denen die verschiedenen Systeme arbeiten, sind zu unterschiedlich, als dass einheitliche Bewertungskriterien festgelegt werden könnten. Besonders im Bereich Robotersysteme existieren nur wenige sprachverstehende Systeme, die mit dem hier vorgestellten Ansatz vergleichbar sind (siehe Kapitel 4.4). Die Unterschiede der einzelnen Systeme machen sich vor allem im Bereich des Anwendungsszenarios bemerkbar, das Aspekte wie Sprachumfang, Echtzeitfähigkeit, Satzkonstellationen und Eingebundenheit in ein übergeordnetes System beinhaltet.

Erschwerend kommt hinzu, dass es sich in dem hier vorliegenden Ansatz um die Evaluation einer Komponente handelt, die in ein Gesamtsystem integriert ist (in diesem Fall das vollständige Dialogverhalten des Roboters). Daher wird in der Evaluation nicht die Komponente allein bewertet, sondern auch die am gesamten Prozess beteiligten anderen Komponenten. Untersucht man das Gesamtverhalten des Robotersystems, so kann oft nicht vollständig geklärt werden, an welcher Stelle sich Schwachstellen und Engpässe im Prozess befinden, denn ein Misslingen der Kommunikation kann von einem einzelnen Modul ausgelöst werden und in Folge wird das Gesamtsystem schlecht bewertet. Evaluiert man dagegen nur ein einziges Modul, können nur Teilaspekte bewertet werden. Zum Beispiel kann die Interaktion über die Schnittstellen zu anderen Modulen nur teilweise berücksichtigt werden. Auch kann nur schwer geklärt werden, wie gut das System die von anderen Modulen erhaltenen Daten verarbeiten kann und umgekehrt. Eine Evaluation kann daher immer nur einen Anhaltspunkt für die Qualität der zu untersuchenden Komponente bieten.

Die Bewertung des gesamten Konzeptes findet daher unter verschiedenen Blickwinkeln statt. Zunächst werden die Fähigkeiten und Grenzen des Mechanismus an sich bewertet. Hier wird untersucht, welche Art Sätze das System verarbeiten kann und welche nicht mehr korrekt analysiert werden können (Kap. 10.1). Dabei gehen auch die Varianten der Verarbeitung in die Analyse mit ein. Danach wird in Abschnitt 10.2 die Fähigkeit der Analyse der situierten Spontansprache aus dem Hometour-Korpus untersucht, der in Kapitel 6.3 beschrieben wurde. Um den Einsatz des sprachverstehenden Systems in einem Roboter-Gefährten bewerten zu können, wird in Abschnitt 10.3 das System während des laufenden Betriebes im Robotersystem BIRON evaluiert. Dort wird ebenfalls die Leistung innerhalb der bestehenden Plattform insgesamt betrachtet. Schließlich wird das Konzept des *robusten Sprachverstehens mit situierten semantischen Einheiten* im Vergleich zu klassischen Analysemethoden bewertet (Kap. 10.4). In Tabelle 10.1 werden die verschiedenen Evaluationsbereiche dargestellt.

Merkmale / Funktion	Daten	Online / Offline	Qualitative / Quantitative Analyse
Fähigkeiten allgemein (10.1)	artifiziell und natürlich	offline	qualitativ
Verarbeitung von Spontansprache (10.2)	natürlich	offline	qualitativ
Gesamtfunktionalität im Robotersystem (10.3)	natürlich	online	quantitativ
Vergleich mit klassischen Ansätzen (10.4)	natürlich	online	qualitativ und quantitativ

Tabelle 10.1.: Die Evaluationen nach verschiedenen Kriterien klassifiziert

10.1. Fähigkeiten und Grenzen des Verstehensprozesses

In diesem Abschnitt werden zunächst einmal die Fähigkeiten und die Grenzen des Verarbeitungsmechanismus allgemein beschrieben. Hier wird geklärt, welche Äußerungen vollständig verarbeitet werden können und welche nur in Ansätzen. Zusätzlich wird berücksichtigt, dass die Verarbeitung innerhalb eines echtzeitfähigen Systems eingesetzt wird. Mit einer vollständigen Suche können alle möglichen Interpretationsvarianten einer Äußerung gefunden werden. Dann jedoch wächst der Suchraum exponentiell und übersteigt womöglich die Rechenzeit, die bei der direkten Interaktion mit einer Person vertretbar ist. Daher muss hier zwischen Rechenzeit und dem Finden aller Varianten abgewogen werden. Der Einsatz von Heuristiken stellt dabei eine gute Möglichkeit dar. Setzt man eine Heuristik ein, um die Verarbeitungszeit sinnvoll einzuschränken, kann entweder eine komplexe Analysemethode verwendet werden oder die Auswahl der Ergebnisse durch einen ausgefeilten Bewertungsmechanismus gesteuert werden. Beides wurde in der hier vorliegenden Arbeit in gewissen Grenzen eingesetzt (siehe Kap. 9.2.3 und Kap. 9.2.4), jedoch mit dem Ziel, den gesamten Mechanismus so einfach wie möglich zu gestalten.

Zunächst wird beschrieben, mit welcher Heuristik welche Äußerungen verarbeitet werden. Die Analyse wurde mit dem Explorations-Werkzeug, unter Verwendung des englischen Lexikons (siehe Kap. 8.1), vorgenommen. Daher werden die meisten Äußerungen auch auf Englisch beschrieben. In einigen Fällen war es jedoch sinnvoller, sie auf Deutsch darzustellen. Insgesamt sind die Äußerungen eher synthetisch, um die Fähigkeiten und Grenzen des Systems aufzuzeigen. Die Evaluation der tatsächlich geäußerten Sätze ist in den nachfolgenden Abschnitten beschrieben.

10.1.1. Funktionsfähigkeit der Heuristikvarianten

In diesem Abschnitt werden die in Kapitel 9.2.4 beschriebenen Heuristiken vorgestellt. Jedoch stellen nur die ersten beiden Strategien (Links-Rechts-Suche und Proximal-Suche) aufgrund der Einfachheit und der Mächtigkeit die zentralen Ansätze dar und daher wird auch die Analyse auf diese beiden Ansätze beschränkt.

Eine sehr einfache Analyse-Strategie ist die Suche von links nach rechts. Sie wurde im System während der Experimente mit dem Roboter BIRON eingesetzt. Es zeigt sich, dass für die in der Hometour vorkommenden spontansprachlichen Äußerungen diese Methode bereits ausreicht, um gute Ergebnisse zu liefern (vgl. Kap. 10.3). Äußerungen, in denen mehrere Wörter vorkommen, die semantisch mit verschiedenen anderen Wörtern kombiniert werden können, sind mit dieser einfachen Heuristik jedoch nicht mehr zu verarbeiten. Ein Beispiel ist die Äußerung „the flower and the big plant“ (die Blume und die große Pflanze).[1] Hier würde der Mechanismus zuerst die SSU Object_plant mit allen möglichen Informationen verbinden und somit das Attribut fälschlicherweise mit dem ersten Objekt verlinken. Mit der Proximal-Suche, die die Nähe der Wörter als ausschlaggebendes Kriterium für den Verlinkungsprozess nutzt, kann diese Art der Äußerungen jedoch semantisch korrekt interpretiert werden. Das Ableitungsergebnis ist in Abbildung 10.1 dargestellt. Genauso können auch Äußerungen wie „can you see the flower and the big plant“ analysiert werden (siehe Abb. 10.2).

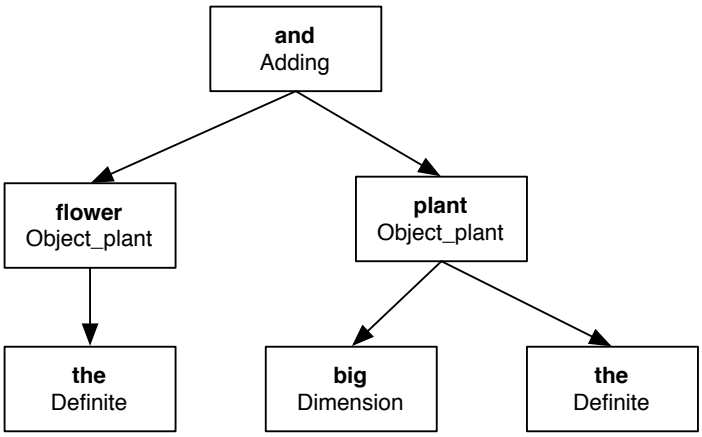


Abbildung 10.1.: Die Ableitungsstruktur der Äußerung „the flower and the big plant“ nach dem Ansatz mit Proximal-Suche

Mit der Proximal-Suche können nicht nur die Attribute jeweils zu den Objekten korrekt zugeordnet werden, sondern auch die Beziehung der Objekte zueinander korrekt dargestellt werden. In der Äußerung „the red can between the yellow can and the blue cup“ (siehe Abb. 10.3) wird das Wort „between“ (zwischen) auf die SSU *In_between* referenziert.

[1] Das Beispiel ist eher theoretischer Natur, da solche Art Sätze in den Äußerungen der Experimente nicht vorgekommen sind.

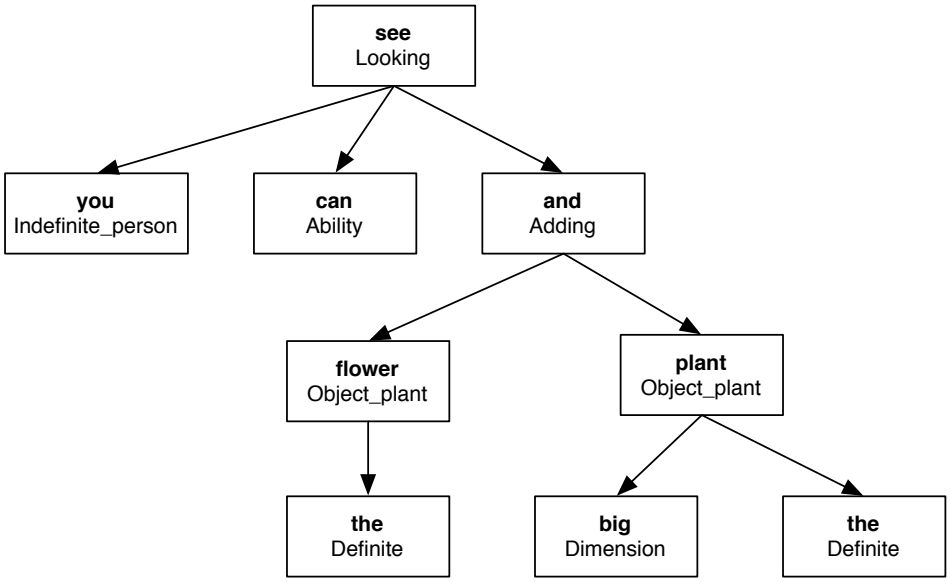


Abbildung 10.2.: Die Ableitungsstruktur der Äußerung „can you see the flower and the big plant“ nach dem Ansatz mit Proximal-Suche

Wie noch in weiteren Beispielen der nachfolgenden Abschnitte gezeigt wird, sind die hier vorgestellten Heuristiken der Links-Rechts-Suche und die Proximal-Suche nicht nur für die hier vorgestellten Satzkonstrukte, sondern auch für andere Äußerungsarten geeignet, wie in nachfolgenden Beispielen dargestellt. In diesem Abschnitt jedoch sollte zunächst ein erster Eindruck über die generelle Funktionsweise der Verarbeitungsstrategie vermittelt werden. Mit den Heuristiken lassen sich die meisten in den Experimenten geäußerten Aussagen verarbeiten. Wie in Kapitel 10.3 gezeigt wird, ist die Links-Rechts-Strategie aufgrund der Einfachheit und Effizienz für den Einsatz in Robotersystemen wie ihn der Roboter BIRON darstellt, bestens geeignet. Sollen auch etwas komplexere Äußerungen verarbeitet werden können, so eignet sich die Proximal-Suche, wie in den folgenden Abschnitten an verschiedenen Beispielen gezeigt wird.

10.1.2. Verarbeitung von Homonymen

Mit dem heuristischen Verfahren der Proximal-Suche ebenso wie mit der Links-Rechts-Suche können auch Homonyme problemlos verarbeitet werden, wie beispielsweise die Äußerung „can you see the can“ (siehe Abb. 10.4) mit dem Homonym „can“. Dabei kann das Wort „can“ sowohl ein Nomen (SSU *Object*) als auch ein Hilfsverb (SSU *Ability*) repräsentieren. Ebenfalls lassen sich auch die etwas komplexeren Äußerungen „the red can between the yellow can and the blue cup“ (siehe Abb. 10.3) oder „can you see the red can between the yellow can and the blue cup – can you see it“ mit diesem Ansatz korrekt interpretieren. Die spontansprachliche Form der letzteren Äußerung stellt hierbei kein Problem dar. Diese Form der Äußerungen werden im

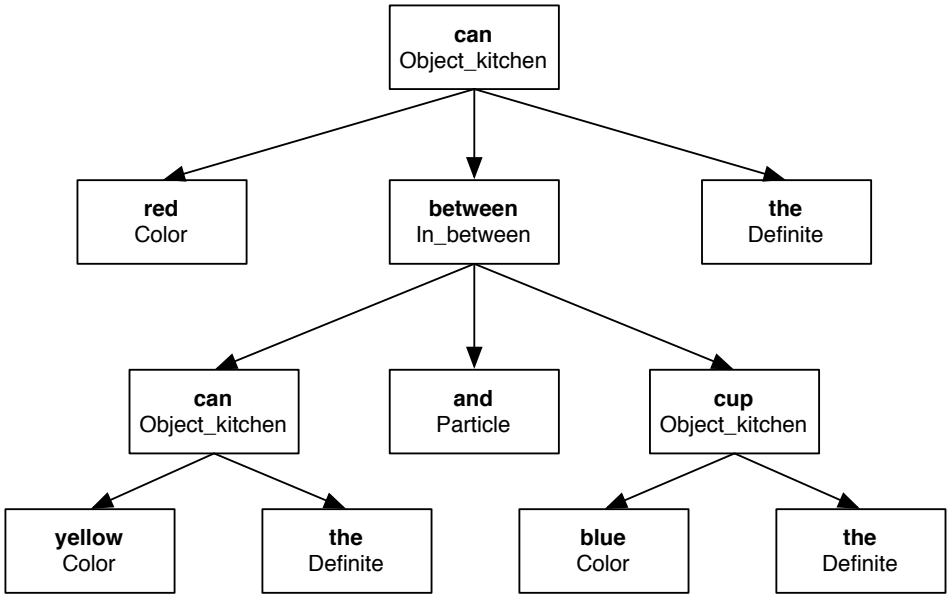


Abbildung 10.3.: Die Ableitungsstruktur der Äußerung „the red can between the yellow can and the blue cup“ mit dem Homonym „can“ nach dem Verfahren der Proximal-Suche

Rahmen des Hometour-Szenarios in Kapitel 10.2 ausführlicher behandelt. Bei der Verarbeitung von Homonymen lassen sich längere Verarbeitungszeiten nicht immer vermeiden, da der Parsemechanismus für jede Homonym-Kombination jeweils vollständig durchlaufen werden muss. Die Rechenzeit verdoppelt sich dadurch bei jedem weiteren Homonym. Insgesamt zeigt sich also ein exponentielles Laufzeitverhalten bezogen auf die Anzahl der Homonyme, was im realen Einsatz im Robotersystem für übliche Längen von Äußerungen jedoch unproblematisch ist (vgl. Kap. 10.3). Insgesamt enthält das Lexikon 26 Homonyme, die im Anhang A.3 abgebildet sind.

10.1.3. Verarbeitung von anaphorischen Äußerungen

Die Äußerung „water the flower – the one next to the door“ kann je nach Anwendungssituation unterschiedlich verarbeitet werden. Existiert die Anaphernauflösung, so kann das Wort „one“ auf die SSU *Object_anaphoric* referenzieren. In diesem Fall entstehen als Ergebnis zwei Teilbäume (siehe Abb. 10.5), die beide an den Dialogmanager weitergeleitet werden. Um entscheiden zu können, ob sie jeweils korrekt erstellt wurden und nicht womöglich eine Fehl-Erkennung zu diesem Ergebnis geführt hat, kann die Vollständigkeit der einzelnen Teilbäume, also die Anzahl der verbundenen und offen gebliebenen obligatorischen Links zu den in Relation stehenden SSUs, ein Bewertungskriterium darstellen. Zusätzlich kann die Reihenfolge der Wörter ein weiteres

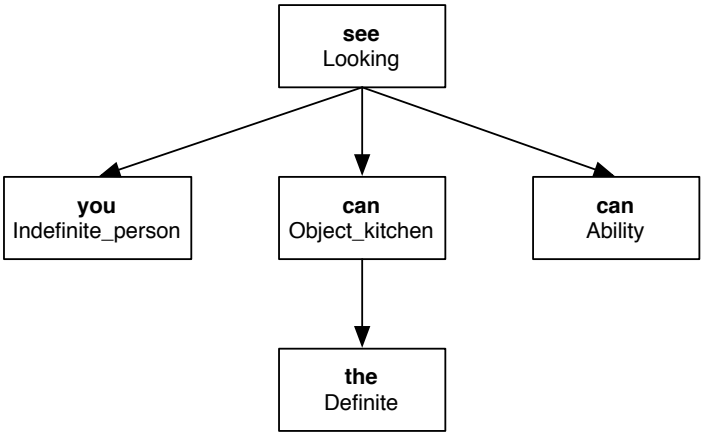


Abbildung 10.4.: Die Ableitungsstruktur der Äußerung „can you see the can“ mit dem Homonym „can“ nach dem Ansatz der Proximal-Suche

Entscheidungsmerkmal bieten. Wenn die beiden Teilbäume jeweils aus linear nebeneinanderstehenden Wörtern entstanden sind, also keine Überlappungen enthalten, so kann mit großer Wahrscheinlichkeit davon ausgegangen werden, dass es sich um zwei einzelne Einheiten handelt. Ansonsten lag vermutlich eher ein Erkennungsfehler vor. Die Einheit „water the red – hello biron – flower“ die zu zwei Teilbäumen mit überlappender Struktur führt, wurde mit nur sehr geringer Wahrscheinlichkeit so geäußert. Dagegen können für die Aussage „biron look – funny – do you like it“ mehrere voneinander strikt getrennte Teilbäume generiert werden. Hier wurden im Vorfeld die einzelnen Äußerungsteile nicht richtig voneinander getrennt, und so an das Sprachverstehen weitergereicht, wie es auch vereinzelt im realen Betrieb vorgekommen ist.[2] Für das Deutsche gilt dasselbe. In der Äußerung „gieße die Blume – die rote“ wird das Personalpronomen „die“ auf die SSU *Object_anaphoric* referenziert.

Im aktuellen Dialogsystem des Roboters BIRON werden jedoch die Anaphern im System nicht aufgelöst, daher wird auch die SSU *Object_anaphoric* nicht verwendet. Für diesen Kontext liefert das System jedoch ebenfalls weitestgehend sinnvolle Interpretationen. Dann wird ein einzelner Baum als Ergebnis geliefert, mit einzelnen alleine stehenden Wörtern „the“ und „one“, die nicht in die Ableitung integriert wurden. Existiert im System keine Anaphernauflösung, können Äußerungen wie „the blue cup and the red one“ nicht korrekt aufgelöst werden. Dann wird nur der erste Teil richtig analysiert, dagegen der Teil „the red one“ nicht integriert. Solche anaphorischen Äußerungen kamen jedoch in den in dieser Arbeit vorgestellten Korpora nur selten vor und wurden daher bisher im Dialogsystem nicht berücksichtigt. Es ist aber bereits eine Erweiterung des Systems um Anaphernresolutionen geplant.

[2] In der auf dem Robotersystem BIRON installierten Version wird immer nur ein Ergebnis weitergereicht, was für die bisherigen Dialoge durchaus ausreichte. Zukünftig bietet dieser Ansatz jedoch eine sinnvolle Erweiterung.

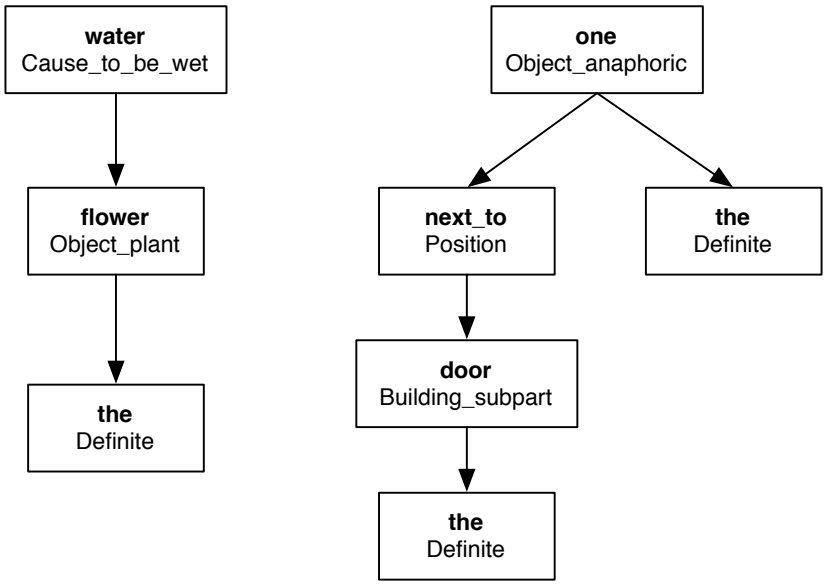


Abbildung 10.5.: Die Ableitungsstruktur der Äußerung „water the flower – the one next to the door“ mit der SSU *Object_anaphoric*

10.1.4. Verarbeitung von Spontansprache

Die meisten spontansprachlichen Äußerungen können ebenfalls verarbeitet werden. Hier soll jedoch nur ein grober Überblick über die Fähigkeiten des Systems gegeben werden, die Verarbeitung der spontansprachlichen Äußerungen aus den Experimenten mit dem Robotersystem BIRON werden in Kapitel 10.2 ausführlicher erläutert.

Ein in der Hometour häufig auftretendes Phänomen sind Wortwiederholungen, wie z. B. in „look here – here“ oder „this is a cup – a cup“. Sie können mit dem in dieser Arbeit beschriebenen Ansatz recht gut verarbeitet werden. Dabei wird nur eins der doppelten Wörter in das Ergebnis integriert, das jeweils andere doppelte Wort wird als einzelne separate Struktur angesehen. Ebenso werden Füllwörter oder unbekannte Wörter in der Verarbeitung ignoriert und die Äußerung so interpretiert, als kämen sie nicht darin vor.

Abbrüche werden soweit verarbeitet, wie sie semantisch in Beziehung stehende Informationen beinhalten. Die Äußerung „can you see the ...“ wird so weit interpretiert wie möglich. Bleiben dabei Links zu obligatorischen Konzepten offen, kann dies ebenfalls an den Dialogmanager weitergereicht werden, in diesem Fall der Link der SSU *Looking* auf die SSU *Object*. Damit weiß der Dialogmanager, dass die Objektinformationen fehlen und er kann gezielt nachfragen.

Bei einem Neuansatz wird die Interpretation weitergereicht, die über die größte semantische Kohärenz verfügt. Für die Äußerung „kannst du die rote – gib mir die rote Tasse“ werden beispielsweise zwei separate Bäume erzeugt. In diesem Fall wird der zweite Baum ausgewählt, entsprechend der Äußerung „gib mir die rote Tasse“, da er die meisten Verbindungen besitzt.

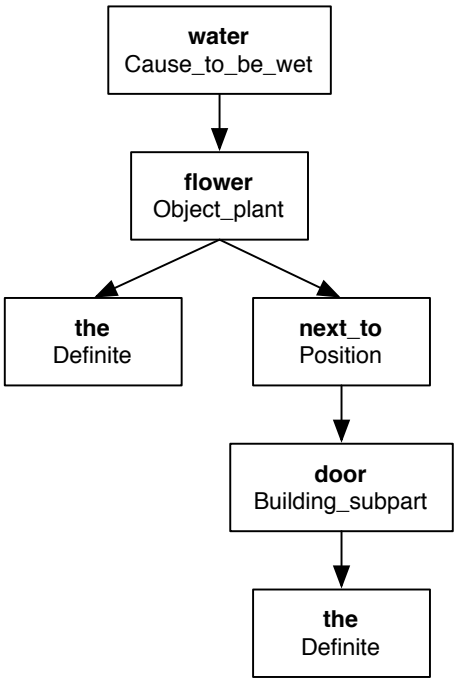


Abbildung 10.6.: Die Ableitungsstruktur der Äußerung „water the flower – the one next to the door“ ohne die SSU *Object_anaphoric*

Selbstkorrekturen können oft integriert werden, jedoch bleiben manche Korrekturen unberücksichtigt. Die Äußerung „die Tasse – die gelbe“ wird, wie bei der Behandlung von Anaphern beschrieben, analysiert. Hier werden bei der Variante mit Anaphernauflösung zwei Ableitungen mit jeweils Teilinformationen generiert, die von der Anaphernauflösung in eine einzige Interpretation zusammengefügt werden muss. Existiert keine Anaphernauflösung, so übernimmt das Sprachverstehen diesen Bereich für die meisten Äußerungen. In diesem Fall wird die Äußerung zu einer einzigen Interpretation überführt, mit einer Ableitung „die gelbe Tasse“ und dem Partikel „die“. Genauso wird die Äußerung „die rote äh die gelbe Blume“ verarbeitet.

Nicht verarbeitet werden können Korrekturen wie „die große Kasse äh Tasse“. In diesem Fall wird das Wort „Kasse“, falls es im Lexikon existiert, mit den anderen Wörtern verbunden. Ebenso schwierig ist „nimm die rote Tasse äh gelbe Tasse“. Hier müsste das System mit berücksichtigen, dass Korrekturen meist die letztgenannte Information ist, anstelle der direkten Nähebeziehung. Dabei wird die Korrektur mit dem Korrektur-Partikel „äh“ angedeutet. Da gerade am Anfang oder am Ende einer Äußerung häufig Informationen abgeschnitten oder falsch erkannt werden (siehe 10.3), könnte ebenso eine Fehl-Erkennung vorliegen. Sie von spontansprachlichen Phänomenen zu unterscheiden, zeigt dem System mitunter seine Grenzen auf.

10.1.5. Verarbeitung von Nebensätzen

Wie auch für die syntaktische Analyse Nebensätze hohe Anforderungen an den Parsingmechanismus stellen, sind Nebensätze für die rein semantische Analyse mitunter schwierig zu interpretieren. In bestimmten Kontexten können Nebensätze ebenfalls verarbeitet werden, jedoch ist der Einsatz des in dieser Arbeit beschriebenen Verfahrens hierfür nur bedingt geeignet. Wie sich in den aufgenommenen Korpora zeigt, kommen Nebensätze nur selten vor.

Bei Infinitiv-, Modal-, Kausalsätzen etc. werden zwei Teilbäume erzeugt, die bei der Bereitstellung geeigneter SSUs (z. B. für „um“, „ob“, „weil“ oder „da“) in einigen Fällen zu einer Einheit zusammengeführt werden können. Dabei repräsentieren die verschiedenen Konjunktionen die Bindeglieder zwischen den Teilbäumen. In vielen Fällen existieren Bindewörter jedoch nicht. Dann können nur die einzelnen Teile für sich an den Dialogmanager weitergereicht werden. Die Aussage „Sie ging nach vorn um besser sehen zu können.“, kann mit diesem Ansatz zu einer einzigen Interpretation zusammengeführt werden, „Er versprach, sich mehr Mühe zu geben.“ jedoch nicht.

Weitere Formen, die vielfach zu zwei voneinander getrennten Teilbäumen führen, sind Relativsätze, Partizipialsätze oder Konditionalsätze. Dabei wird die Analyse mitunter durch Ellipsen noch weiter erschwert. Der Relativsatz „Klaus, der morgen kommt, mag keinen Fisch.“ kann nur zum Teil analysiert werden. Er wird überführt in die Teilstrukturen „Klaus kommt morgen“ und „ mag keinen Fisch“, wobei im zweiten Teil die Person fehlt. Hier müsste sozusagen die Person des ersten Teils automatisch auch der zweiten Teilstruktur zugewiesen werden. Bei dem Satz „Klaus, der Bruder von Nadine, mag keinen Fisch.“ wird je nach Analysevariante entweder die Person „Klaus“ dem zweiten Teil zugeordnet, oder aufgrund der Nähebeziehung „der Bruder von Nadine“.

Insgesamt ist es daher schwierig, solche komplexen Sachverhalte darzustellen und demnach auch, sie zu analysieren. Werden Nebensätze durch ein besonderes Wort wie „um“ markiert, können dafür ebenfalls spezielle SSUs bereitgestellt werden, mit deren Hilfe die Nebensätze mit dem Hauptsatz verbunden werden können. Dennoch wurden für dieses Problem Methoden entwickelt, die hierfür besser geeignet sind, wie z. B. die Grammatiktheorie der HPSG für das Deutsche [Mül99]. Diese mit komplexen syntaktischen Regeln ausgestatteten Ansätze wiederum sind umgekehrt nicht für die Verarbeitung von Spontansprache ausgelegt.

Da neben den spontansprachlichen Phänomenen zusätzlich noch Fehl-Erkennungen auftreten können, existieren für die Analyse dieser besonderen Form von Sätzen weitere Probleme. Es ist alleine schon schwierig, eine Selbstkorrektur wie „Klaus, der morgen kommt – (äh) übermorgen kommt“ von einem Relativsatz wie „Klaus, der morgen kommt – übermorgen wieder geht“ zu unterscheiden. Kommen noch zusätzliche Erkennungsfehler hinzu, steht die Analyse vor besonderen Herausforderungen. Insgesamt stellt das Problem der Nebensätze sowohl für die syntaktische als auch für die semantische Analyse noch eine schwierige Aufgabe dar. Jedoch soll dies nicht Thema der hier vorliegenden Arbeit sein.

10.1.6. Reihenfolgebeziehungen

Ein größeres Problem stellt die Berücksichtigung von Reihenfolge-Informationen dar. Sie werden in einem Graphen und ebenfalls in einem XML-Schema nicht abgebildet. Bei der Aussage „a folgt b“ bestimmt jedoch die Reihenfolge der Wörter die semantische Beziehung der Teile „a“ und „b“. Im Englischen kann davon ausgegangen werden, dass auf der linken Seite das Subjekt oder der Akteur steht und rechts das Objekt. Sie lassen sich in den meisten Fällen ebenfalls durch den Kasus bestimmen („sie“, „ihr“ usw.). Ist eine Unterscheidung möglich, bevorzugen die SSUs ebenfalls bestimmte andere SSUs, die sich mitunter auch auf verschiedene Beugungsfälle zurückführen lassen (siehe Kap. 8.2.3). In seltenen Fällen ist dies jedoch nicht möglich, z. B. in „biron follow sonja“. Hier ist in beiden Fällen *Name_personal* die zugehörige SSU zu den Namen „Sonja“ und „BIRON“. Die Abbildung 10.7 stellt die Ableitungsstruktur der Äußerung dar. Hier bildet die SSU *Move_afterwards* die Wurzel, die offene Linkverbindungen zu den SSUs *Actor* und *Object* besitzt. Diese wiederum können geschlossen werden durch das Wort „BIRON“, das als SSU *Name_personal* auf die Oberkategorie *Actor* abgebildet wird, sowie der SSU *Name_personal* („Sonja“), die ebenfalls der Oberkategorie *Object* angehört. Die SSU *Name_personal* gehört demnach sowohl zur Klasse der Akteure als auch der Objekte.

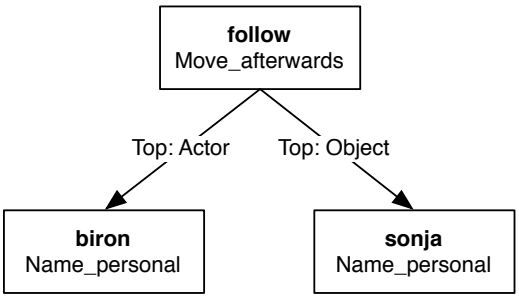


Abbildung 10.7.: Die Ableitungsstruktur der Äußerung „biron follow sonja“ mit den hierarchischen Beziehungen

Ebenso verhält es sich mit bestimmten Fragekonstellationen. Die Frage „Magst du sie?“ wird von der Aussage „Du magst sie.“ nur durch die Reihenfolgebeziehung unterschieden. Jedoch kann dieselbe Äußerung mit einer anderen Betonung („Du magst sie?“) dieselbe Bedeutung bekommen wie die erste. Die Reihenfolge spielt demnach in vielen Fällen keine Bedeutung.

Die Erkennung von Fragen stellt daher insgesamt ein großes Problem dar, denn bestimmte Äußerungen können sowohl Frage als auch Aussage, oder sogar einen Befehl darstellen, wie die Äußerung „du gehst zum Arzt“. Nur an der Betonung läßt sich hier erkennen, wie sie interpretiert werden soll. Daher wird die Eingliederung der Prosodieinformationen in komplexeren Szenarien unumgänglich (vgl. Kap. 8.2.3). Fragen, die durch ein Hilfsverb ausgewiesen werden, wie in „Kannst du mir die Butter geben?“ oder „kannst du mir folgen“ können oftmals als Frage oder als indirekte Aufforderung interpretiert werden. In der letzten Frage wird der Bedeutungsunterschied zu „du kannst mir folgen“ eher durch zusätzliches Kontextwissen als durch die Reihenfolgebeziehung bestimmt.

Für die Interaktion mit dem Roboter scheinen diese zusätzlichen Informationen jedoch nach den Erfahrungen aus den Experimenten nicht notwendig zu sein, da diese Fragekonstellationen eher selten vorkommen und das Robotersystem zur Zeit über keinen Inferenzmechanismus zur Beantwortung von Fragen verfügt.

Zur Zeit wird die Standard-Vorgabe verwendet, die zuerst immer links und dann rechts sucht. Damit lassen sich die Probleme der Erkennung des Akteurs in der Regel lösen. Für Fragen, die mit einem Verb beginnen, könnte man zusätzlich einen einfachen Mechanismus einfügen, der genau dies überprüft und die Äußerung dann insgesamt als Frage markiert[3]. Fragen mit Hilfsverbkonstellation werden nur indirekt markiert, indem die Interpretation eine mögliche Frage durch das Wort „do“ oder can“ indirekt markiert. Hier muss der Dialogmanager je nach Kontext entscheiden, ob er die Äußerung als Anweisung oder als Frage bewertet.

Dennoch ist dieser Ansatz nur zum Teil befriedigend. Eine weitere Idee ist, die Reihenfolge-Informationen innerhalb der SSUs mit abzuspeichern, wie in Kapitel 11.2 ausführlicher erläutert wird. Dann hat man sozusagen neben den rein semantischen Informationen lokale syntaktische „weiche“ Regeln. Die Verarbeitungsstrategie könnte dann diese Regeln berücksichtigen. Dennoch sollen sie keine festen Vorgaben darstellen, sondern nur die Präferenz angeben, so dass auch Abweichungen aufgrund von Spontansprache verarbeitet werden können. Es können demnach Objekteigenschaften darin vermerkt sein oder dass der Akteur nach links beginnend zu suchen ist. Dabei wird im Gegensatz zu globalen syntaktischen Regeln die lokale Sicht der SSU angenommen. Aus den Erfahrungen bei der Erstellung der SSUs zeigte sich, dass diese Angaben durchaus in vielen Fällen machbar sind. Es zeigte sich, dass bei:

- *Aktionen* der zugehörige Akteur eher links zu finden ist, beteiligte Objekte eher rechts. Beispiel: „ich gieße die Palme“; Akteur steht links, das Objekt auf der rechten Seite. In seltenen Fällen der Passivkonstruktion gilt diese Reihenfolge jedoch nicht.
- *Objekten* die zugehörigen Informationen eher links zu finden sind. Beispiel: „die rote Tasse“; Artikel und Attribut stehen links.
- *Positionen* die zugehörigen Informationen meist rechts stehen. Beispiele: „neben der Tür“ oder „auf dem Tisch“; das in Beziehung stehende Objekt befindet sich rechts.
- *Konjunktionen* die Informationen sowohl links als auch rechts zu finden sind. Beispiel: „Gerburg und Steffi“; die Eigennamen umschließen die Konjunktion räumlich.

Die hier vorgeschlagene Erweiterung bietet vor allem den Vorteil, gezielt nach bestimmten Informationen suchen zu können. Dann kann z. B. der Folger vom Gefolgten aufgrund der Reihenfolgenbeziehung erkannt werden. Semantische und syntaktische Informationen können mit dieser Erweiterung gleichzeitig berücksichtigt werden. Nachteilig wäre, dass die SSUs wesentlich aufwendiger zu erstellen sind und dafür zusätzlich mehr linguistische Kenntnisse erforderlich sind.

[3] Hier muss jedoch genauer geprüft werden, um sicherzustellen, dass es sich nicht um einen Satz im Imperativ oder Konjunktiv handelt.

Ebenfalls müsste die Frage geklärt werden, wieviele zusätzliche Regeln den Prozess unterstützen und welche unnötig sind (oder sogar störend). Ein alternativer Ansatz zur robusten Verarbeitung von Sprache ist die Verwendung von gewichteten Constraints. Dieses Verfahren wurde in der Dependenzgrammatik verwendet [Men02], um Äußerungen ausdrücken zu können, die nicht vollständig wohlgeformt sind.

Für die Erkennung von indirekten Fragen und zur Unterscheidung von Fragen und indirekten Anweisungen, lösen die oben genannten Ansätze diese Problematik wie erläutert nur zum Teil. Soll der Roboter über komplexe Interaktionsfähigkeiten verfügen, sind Prosodieinformationen (neben einem Inferenzmechanismus) ein weiterer wichtiger Bestandteil für den gesamten Dialogverlauf.

10.1.7. Fazit

Die in diesem Abschnitt durchgeführte Evaluation diente der qualitativen Bewertung der entwickelten Datenstrukturen zusammen mit den wichtigsten heuristischen Verarbeitungsstrategien. Es zeigt sich, dass sowohl die Verarbeitung nach der Heuristik der Links-Rechts-Strategie, aber insbesondere die Proximal-Suche mit nur sehr wenig Aufwand gute Ergebnisse für ein breites Spektrum an Äußerungen liefert. Das Verfahren erzeugt aus einer großen Menge von Äußerungen sinnvolle Ableitungsstrukturen, die für die Gewinnung semantischer Informationen wesentlich sind. Es hat sich insgesamt gezeigt, dass

- **Homonyme** in den Verarbeitungsprozess integriert werden können.
- **Äußerungen**, in denen Anaphern vorkommen, verarbeitet werden können. Hier unterstützt das System je nach Fähigkeit den gesamten Dialog. Existiert eine separate Komponente zur Anaphernresolution, so kann es die dafür erforderlichen Informationen erzeugen und bereitstellen. Existiert jedoch kein eigenständiges Modul für die Auflösung von Anaphern, so übernimmt das System diese Aufgabe so weit wie möglich selbst.
- **Spontansprache** gut mit diesem Ansatz verarbeitet werden kann. Der Mechanismus kann Äußerungen auch interpretieren, wenn spontansprachliche Phänomene darin vorkommen. Es kann Wortwiederholungen, Satzabbrüche, Neuansatz und Selbstkorrekturen verarbeiten. Einzig schwierig sind Korrekturen, bei denen das verbesserte Wort weiter rechts steht. Dieses ist in vielen Fällen weder mit der Links-Rechts-Strategie noch mit dem proximalen Ansatz aufzulösen. Zusätzlich ist diese Form von Korrekturen nur schwer von Spracherkennungsfehlern zu unterscheiden. Eine mögliche Lösung könnte hier die Berücksichtigung des Korrektursignals „äh“ sein, das markiert, dass daran anschließend die korrekte Form folgt.
- **Nebensätze** so weit zu verarbeiten sind, wie besondere Wörter wie z. B „da„ „weil“ oder „um“ diese Nebensatzkonstruktion signalisieren. Andere Nebensatzkonstruktionen sind

nur schwer zu verarbeiten. Erschwerend kommt hinzu, dass sie von spontansprachlichen Phänomenen kaum zu unterscheiden sind.

- **Reihenfolgebeziehungen** ebenfalls problematisch sein können. Hier können Informationen der Richtungspräferenz auf lokaler Ebene der SSUs den Prozess unterstützen.

Im Gegensatz zu dem hier vorgestellten Ansatz mit quadratischer Laufzeit würde das Finden aller Lösungen für die Verarbeitungszeit einen exponentiellen Anstieg bedeuten. Alternativ könnte man verschiedene Heuristiken nacheinander einsetzen und dann das insgesamt beste Ergebnis übernehmen. Dabei verlängert sich die Rechenzeit entsprechend. Ebenso ist es möglich, komplexere Mechanismen einzusetzen, wie die Integration der Reihenfolge-Informationen. Jedoch muss hier zwischen Aufwandskosten und dem Nutzen für die jeweilige Anwendung abgewogen werden. Für die Analyse der Äußerungen aus den Experimenten zur Erstellung des Hometour-Korpus reicht bereits die einfache aber auch schnelle Suche von links nach rechts aus, um gute Ergebnisse zu liefern. Hier waren eher die von den Probanden geäußerten unbekannten Wörter das begrenzende Element, als der Verarbeitungsmechanismus (vgl. Kap. 10.3) selbst.

10.2. Verarbeitung der Spontansprache aus dem Hometour-Korpus

In diesem Abschnitt wird die qualitative Analyse von Äußerungen aus einem realen Experimentsetting beschrieben, die im nächsten Kapitel für die quantitative Analyse verwendet werden. Die Probanden konnten sich dabei ohne Einschränkungen frei äußern und haben damit sozusagen die Grenzen des Sprachverstehens ausgelotet. Es werden Verarbeitungsergebnisse von realen Daten aus dem Experiment-Korpus der Hometour ausgewählt (vgl. Kap. 6.2) und ausführlicher beschrieben [Hüw06b]. Insbesondere werden in diesem Abschnitt markante spontansprachliche Äußerungen, die die Probanden während des Experiments äußerten, diskutiert. In Abbildung 10.8 sind die aufeinanderfolgenden Äußerungen der Probanden dargestellt. Pausen innerhalb der Äußerungen sind durch „–“ markiert und sprachbegleitende deiktische Gesten durch „< ... >“. Für die Erläuterung der Verarbeitung wird angenommen, dass der Spracherkenner die Äußerung korrekt erkannt hat und diese an das Sprachverstehen weiterleitet.

Die erste Äußerung von *Person1* wird durch den Mechanismus vollständig verarbeitet. Der Roboter BIRON bekommt die Anweisung, ein Objekt anzuschauen und den Hinweis auf eine mögliche Geste: „look – this – is a cube – this – BIRON“. Diese Äußerung enthält eine Wortwiederholung „this“, die während des Verarbeitungsprozesses ignoriert wird. Sie kann beim zweiten Mal nicht mehr in die erzeugte Struktur integriert werden, da keine SSU dafür einen weiteren Platz bereithält. Daher wird die Interpretation etwas schlechter bewertet, als wenn sie ohne Wortwiederholung geäußert worden wäre, jedoch immer noch als vollständig interpretierbar. Die Äußerung „robot look – do you see“ enthält zwei Teile, die hier vom Spracherkenner als ein einzige weitergereicht wurde. Sie wird daher als eine Anweisung interpretiert, dass der Roboter schauen soll (SSU *Looking*). In der Äußerung fehlt die Informationen auf ein Objekt, die als obligatorische

```
Person1: Look - < this - is  a cube - this > - BIRON.
         Robot < look > - do you see?
         < This - is a cow > - funny.
         Do you like it? ...
Person2: And look < here > - a cup - a cup.
         < Look is a cup - cup >.
         So good - < this is a cup >.
         We are getting somewhat. ...
Person3: Please stop here.
         Stop moving - fine.
         What is your name?
         How old are you?
         Have you got any hobbies? ...
Person4: What can you do?
         Oh - so what is that? ...
```

Abbildung 10.8.: Auszüge von Äußerungen während der Experimente des Hometour-Settings

Verlinkung der SSU vermerkt ist. Da nur ein Teil der Äußerung interpretiert werden kann, wird diese Interpretation als *partiell korrekt* vermerkt und der Dialogmanager kann mit der Suche nach einem entsprechenden Objekt oder einer Rückfrage reagieren. Die zweite Äußerung „this is a cow – funny“ wird als eine Objektbeschreibung interpretiert. Hier sind alle wichtigen Informationen vorhanden. Das Wort *funny* (lustig) war zu dem Zeitpunkt der Experimente noch nicht im Lexikon, wurde demnach als unbekannt im Prozess markiert. Da aber alle relevanten Informationen vorhanden sind, wird die Interpretation ansonsten als vollständig angesehen. Mittlerweile stellen Beschreibungen wie „funny“ ein zusätzliches Attribut einer Beschreibung dar. Die Äußerung „do you like it“ wird zur Zeit als Zustandsbeschreibung interpretiert, die jedoch mit dem Wort „do“ eine Hilfskonstellation markiert (vgl. Kap. 10.1.6). Möglich wäre jedoch auch, dass

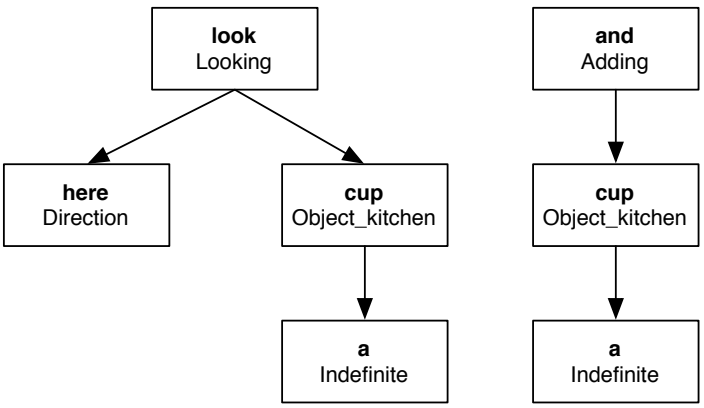


Abbildung 10.9.: Ableitungstruktur der Äußerung „and – look here a cup – a cup“

es gleichfalls als Beschreibung einer Präferenz (*mögen*) interpretiert werden kann, die durch das „do“ (mit dem Link auf „like“) angekündigt wird. Wie in Kapitel 10.1.6 und 8.3 beschrieben, sind Fragen generell schwierig zu erkennen und hierfür wäre die Zuhilfenahme von Prosodie-Informationen (oder in diesem Fall auch syntaktische Informationen) hilfreich, um die genaue Äußerungsbedeutung zu ermitteln. Das Pronomen „it“ erhält einen Hinweis auf eine anaphorische Bedeutung.

Die Äußerungen von *Person2* werden in ähnlicher Weise verarbeitet. Zuerst bekommt der Roboter eine Anweisung, nach einem Objekt zu schauen, genauer nach einem Becher („and look here – a cup a cup“). Der Roboter bekommt ebenfalls Hinweise auf eine wahrscheinliche Geste. Die Informationen aus der Szene und der Sprache können so verbunden und zusammen gespeichert werden. In dieser Ableitung wird wiederum die Wiederholung ignoriert. Dabei werden zwei Teil-Ergebnisse ermittelt. Das erste besteht aus „and a cup“ und das zweite aus „look here a cup“ (siehe Abb. 10.9). Da das zweite besser bewertet wird, wird dieses als Ergebnis ausgegeben, jedoch mit dem Hinweis, dass die Äußerung nur teilweise interpretiert wurde. Jetzt kann der Dialog entsprechend reagieren. In diesem Fall fragt das System nach. Die Antwort „look is a cup - cup“ bewertet das Sprachverstehen als eine vollständige Interpretation, wobei das zweite „cup“ wiederum ignoriert wird. Nachdem der Roboter mit Hilfe der Richtung der Geste das Objekt gefunden hat, gibt der Roboter Bescheid, dass er die gesuchte Tasse gefunden hat. Ist eine Pause zu lang, werden zwei Äußerungen erkannt und jeweils separat verarbeitet. Hier werden dann jedoch die einzelnen Phrasen in der Historie solange gesammelt, bis die Äußerung durch die Spracherkennung als vollständig markiert wird. Bei einer einzigen Äußerung werden diese Informationen gleich mit einer einzigen Interpretation an den Dialogmanager weitergereicht. Die Äußerung „so good – this – is a cup“ wurde auf eine Objektbeschreibung abgebildet. Dabei wurde der Kommentar „so good“ ignoriert. In dem Design des gesamten Robotersystems wurden zunächst Meta-Kommentare, die die Benutzer äußern können, nicht in der Sprachverarbeitung berücksichtigt, hier konnten sowohl der Spracherkenner als auch das Sprachverstehen die Äußerung nicht korrekt erkennen. Dies wurde mittlerweile verbessert, indem die SSUs um einige semantische Konzepte über Meta-Kommentare erweitert worden sind. Jedoch ist es generell schwierig, den Wortschatz soweit zu ergänzen, dass alle möglichen Kommentare verstanden werden können. In einer realen Kommunikationssituation, die den Benutzern auch eine freie Kommunikation erlaubt, kann man nie ausschließen, dass unbekannte Inhalte geäußert werden. Die letzte Äußerung „we are getting somewhat“ wurde daher zu dem Zeitpunkt interpretiert als eine nur teilweise verständliche Äußerung. Mittlerweile entspricht die Interpretation der Äußerung „we get somewhat“ (in Abb. 10.10 dargestellt), da in diesem Roboter-Kontext nicht zwischen *Simple Present* und *Present Progressive* unterschieden wird.

Von *Person3* erhält der Roboter die Aufforderung anzuhalten („please stop here“). Hier bildet die SSU *Process_stop* die Wurzel, die Verweise auf eine SSU *Social_interaction* und auf eine SSU *Direction* besitzt. Die Äußerung „stop moving – fine“ enthält wiederum die SSU *Process_stop* und zusätzlich eine SSU *Move_to* auf die Handlung „moving“. Der Roboter hält daraufhin an. Das Wort „fine“ wird als Meta-Kommentar auch in diesem Fall ignoriert, da es nicht im Lexikon enthalten ist. Die zweite Äußerung „what is your name“ (siehe Abb. 10.11) wird zur Frage nach dem Namen umgewandelt. Hier antwortet der Roboter mit „my name is BIRON“. Die Äußerung

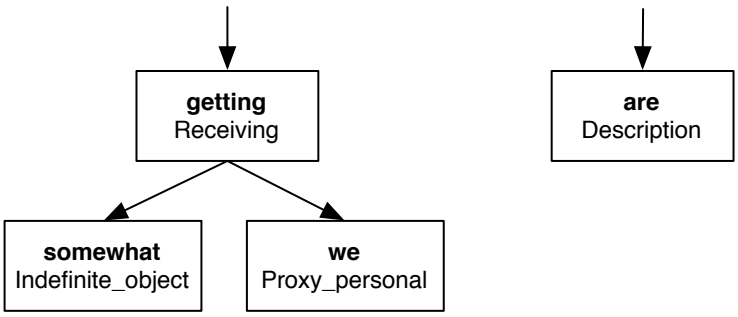


Abbildung 10.10.: Ableitungstruktur der Äußerung „We are getting somewhat.“

„how old are you“ wird interpretiert als eine Frage (SSU *Question_attribute*) nach dem Attribut „old“ der Person „you“, also dem Alter des Roboters. Die letzte Äußerung, die Frage nach den Hobbys („have you got any hobbies“), wird nur zum Teil verstanden, da das Wort „hobbies“ unbekannt ist. Daher wird die Äußerung als eine *Description* verstanden mit der SSU *Possession* als Wurzel und „you“ als SSU *Owner* (also „have you“). Hier fehlen jedoch die Informationen, was besessen wird. Ist „hobby“ (SSU *Object_abstract*) im Lexikon aufgenommen, wird diese Information ebenfalls verlinkt und die Äußerung kann vollständig interpretiert werden.

Bei *Person4* wird die erste Äußerung „what can you do“ als Frage an den Roboter interpretiert. Daraufhin antwortet der Roboter BIRON mit: „I can move to another location, I can follow you and if you show me something, I can remember it“. Die Äußerung „so what is that“ (siehe Abb. 10.12) wird als Frage nach einem Objekt verstanden, die ebenfalls einen Hinweis auf eine Geste enthält.

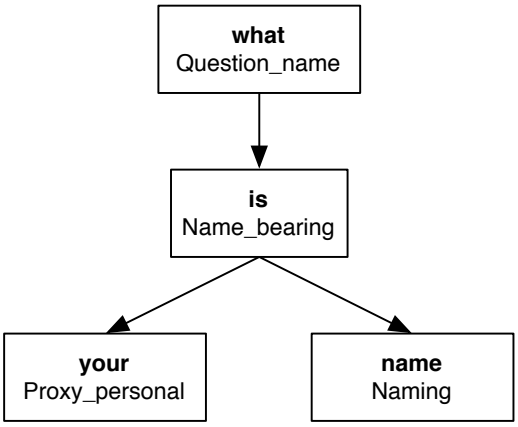


Abbildung 10.11.: Ableitungstruktur der Äußerung „what is your name“

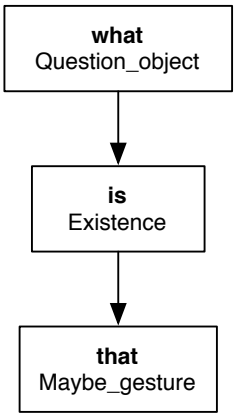


Abbildung 10.12.: Ableitungstruktur der Äußerung „what is that“

10.3. Auswertung der Sprachverstehenskomponente auf dem Robotersystem BIRON

Für die Evaluation des Gesamtsystems wurde die Interaktion von 14 naiven Probanden, im Alter zwischen 12 und 37 Jahren, die bis dahin noch keinen direkten Kontakt mit Robotersystemen hatten, mit dem Robotersystem BIRON analysiert [Hüw06a, Hüw06b]. Die Experimente sollten einerseits zeigen, wie die Funktionalität des Sprachverstehens im Gesamtsystem bewertet werden kann: wie gut die semantischen Analyseergebnisse der spontansprachlichen Äußerungen sind, wie robust das System gegenüber Spracherkennungsfehlern ist und wie die Performanz des sprachverstehenden Systems insgesamt bewertet werden kann. Andererseits sollte insgesamt getestet werden, welche Art Äußerungen Personen in einer uneingeschränkten Kommunikationssituation an einen Roboter richten. Die genaue Beschreibung der Dialoge ist in Kapitel 6.3 festgehalten. Teile der Äußerungen wurden bereits im vorangegangenen Kapitel für die qualitative Analyse verwendet. In diesem Abschnitt werden die Daten quantitativ ausgewertet.

10.3.1. Durchführung des Experiments

Um eine möglichst offene Kommunikationssituation zu schaffen, sollten sich die Probanden im ersten Experimentlauf mit dem System vertraut machen, ohne dass sie weitere Informationen über das System bekamen. Im zweiten Durchlauf erhielten sie dann mehr Informationen über die Funktionalität des Roboters BIRON und die technischen Details (z. B. über Sprachumfang und Fähigkeiten). Die gesamten Experimente verliefen auf Englisch, wobei die Probanden größtenteils Deutsch als Muttersprache besaßen [Li06b]. Aufgrund der großen Freiheitsgrade der Probanden und der Unkenntnis über die genaue Funktionalität und den Wortschatz, den der Spracherkenner verarbeiten kann, lag die Fehl-Erkennungsrate des Spracherkenners (pro Äußerung) im ersten Lauf bei 60% und im zweiten Lauf bei 42%, insgesamt im Durchschnitt also bei 52%.

Wie in Kapitel 6.3 beschrieben äußerten die Probanden tendenziell kurze Sätze und machten Pausen innerhalb der Äußerungen. Sie entsprachen in vielen Fällen nicht der Standard-Grammatik und enthielten viele spontansprachliche Phänomene sowie Meta-Kommentare. Auch variierte der Sprachstil der Probanden deutlich. Sie wechselten teilweise in einen kindlichen Sprachstil (engl. *baby-talk*) oder wurden lauter und überartikulierten, was den Erkennungsprozess zusätzlich erschwerte. Typische Äußerungen entsprachen den folgenden Beispielen: „look – cow“ oder „a book this is“. In der freien Interaktion der Probanden mit dem Roboter zeigt sich, wie wichtig es ist, dass das Sprachverstehen gerade auch mit unsicheren Spracherkennungsergebnissen zurechtkommen kann, denn Fehl-Erkennungen können in solch einem Kontext nie ganz verhindert werden.

Das System hat insgesamt 1642 Äußerungen von 14 Probanden während der Experimentphase verarbeitet. Dabei betrug die Interaktionszeit 151 Minuten mit einer durchschnittlichen Interaktionszeit von 5 Minuten und 14 Sekunden je Proband. Das sprachverstehende System hat alle eingegangenen Äußerungen verarbeitet und für jede Äußerung auf einem gängigen PC (Pentium M, 1.4 GHz) im Schnitt 20 Millisekunden Verarbeitungszeit benötigt. Die Anforderung für die Verarbeitungszeit einer Echtzeit-Anwendung wurde somit erfüllt. Die Äußerungen bestanden durchschnittlich aus 3.23 Wörtern, was indiziert, dass die Probanden eher kurze Äußerungen an das Robotersystem richteten. Jedoch auch längere Äußerungen konnten schnell verarbeitet werden. Die längste Äußerung bestand aus 14 Wörtern und wurde innerhalb von 300ms verarbeitet. Auch hier wurde der gesamte Verarbeitungsprozess durch das Sprachverstehen nicht wesentlich beeinträchtigt.

10.3.2. Bewertungsergebnisse

Aufgrund der hohen Fehl-Erkennungsrate war die Bewertung der Ergebnisse aus dem Sprachverstehen ein wichtiges Element für die Interaktionsfähigkeit. Von ihrer Bewertung hing ab, wie das Dialogsystem auf Antworten reagierte. Zu dem Zeitpunkt lag ein sehr einfaches Bewertungssystem zugrunde. Die Anzahl der verlinkten Wörter wurde durch die Anzahl der Wörter in der Äußerung dividiert. Damit konnte ein Wert zwischen Null und Hundert Prozent ermittelt werden. Mittlerweile existieren alternative Berechnungsverfahren, die in Kapitel 9.2.3 beschrieben sind. Wie sich zeigte, reicht jedoch die einfache Berwertung bereits aus, um in diesem Szenario die Güte der semantischen Kohärenz sinnvoll zu bestimmen und somit ein geeignetes Maß für die Handlungsoptionen des Dialogmanagers zu liefern. Äußerungen, die entweder vollständig oder mindestens zu 75% verlinkt werden konnten, wurden mit vollständig korrekt (*fullunderstanding*) bewertet. Äußerungen mit mindestens 50% Kohärenz als teilweise korrekt (*partialunderstanding*) und Äußerungen mit weniger Kohärenz als nicht verständlich (*nonunderstanding*) bewertet. In der aktuellen Version werden jedoch zusätzlich die obligatorischen und optionalen Argumente unterschiedlich stark bewertet, wobei die obligatorischen Informationen stärker in die Bewertung einfließen.

67% der Äußerungen wurden als vollständig korrekt bewertet, wobei davon in 10% der Fälle (absolut 6% Punkte) kein Dialogakt zugeordnet werden konnte und die Äußerung somit vom

Dialogmanager nicht direkt verarbeitet werden konnte. Demnach konnte der Dialogmanager in 61% der Anfragen direkt eine entsprechende Handlung veranlassen oder angemessen antworten. In 39% der Fälle hat der Dialogmanager eine Klärungsfrage an den Benutzer gestellt. Von den 39% der Äußerungen wurden dabei 17% der Äußerungen als zum Teil verständlich (*partialunderstanding*) bewertet. Hierbei konnte die Interpretation genutzt werden, um konkretere Anfragen zu stellen. 22% der Äußerungen wurden als nicht verständlich (*nonunderstanding*) eingestuft. Hier hat der Dialogmanager in der Regel direkt um die Wiederholung des Gesagten gebeten.

Nur 4% der Äußerungen, die korrekt erkannt wurden, wurden vom System falsch interpretiert oder als nicht verständlich abgelehnt. Hier sind die Fehler in den meisten Fällen aufgrund fehlender Einträge im Lexikon entstanden, z. B. durch unbekannte Wörter wie „funny“ oder „glad“, die mittlerweile größtenteils im Lexikon eingetragen sind. In Kapitel 10.4 werden die falsch generierten Interpretationen ausführlicher diskutiert.

Einzig schwierig zu bewerten ist die Generierung der Antworten aus fehlerhaft erkannten Äußerungen. War die Äußerung nur in Teilen fehlerhaft erkannt worden, so konnte bewertet werden, ob der korrekt erkannte Teil auch richtig weiterverarbeitet wurde. Nur in Ausnahmefällen wurde die gesamte Äußerung falsch erkannt, z. B. statt „hello you“ „yellow cube“. Dann kann auch das Sprachverstehen nicht erkennen, dass eine Fehl-Erkennung vorliegt und gibt dieses falsche Ergebnis an den Dialogmanager weiter.

Viele Spracherkennungsfehler entstanden durch das fehlerhafte Anhängen eines zusätzlichen Wortes oder das Abschneiden eines Wortes von der Äußerung. Beispielsweise wurden Äußerungen erkannt wie „what can you do *look*“ oder „this is a cube *let*“. In diesen Fällen verbindet das Sprachverstehen die semantisch zusammengehörigen Informationen und ignoriert das jeweilig angehängte Wort. Fehlt ein Wort in der erkannten Äußerung, so wird die Information, soweit vorhanden, zusammengefügt. Fehlen wesentliche Informationen, so kann das System dies durch offene Links ebenfalls erkennen und eine konkrete Rückfrage kann an den Benutzer gestellt werden.

10.3.3. Schlussfolgerungen

Vergleicht man die Fehlerrate des Spracherkenners mit der Ausgabe des Sprachverstehens, zeigt sich, dass das System insgesamt vom Verarbeitungsprozess profitiert. Lag die Rate der korrekten Erkennung der Äußerungen bei 48%, so konnten doch immer noch in der Mehrzahl der Fälle vollständige Interpretation gewonnen werden. In 61% der Fälle wurde die Äußerung als semantisch korrekt bewertet, in 17% fehlten Informationen, die mit einer konkreten Rückfrage erhalten werden konnten. Nur in 22% wurden die Äußerungen als vollständig falsch verstanden interpretiert. Die Verarbeitung hat demnach nicht nur eine Interpretation der Äußerungen generiert, sondern zusätzlich zwischen falsch erkannten und korrekt erkannten Äußerungen unterschieden. In Tabelle 10.2 ist der Vergleich zwischen der Fehlerrate der Spracherkennung und der Bewertung des Sprachverstehens abgebildet.

Spracherkennung (ASR)		Sprachverstehen (ASU)	
falsch	52%	nicht / teilweise interpretierbar	33%
korrekt	48%	vollständig interpretierbar (* mit Dialogakt Zuordnung)	67% (* 61%)

Tabelle 10.2.: Fehlerrate der Spracherkennung verglichen mit der Bewertung der semantischen Kohärenz des Sprachverstehens

Um Genaueres über die Art der Äußerungen mit unterschiedlicher Bewertung zu erfahren, wurden diese Äußerungen ebenfalls auf ihre Syntax hin analysiert. Mit *grammatikalisch korrekt* wurden Äußerungen bezeichnet, die der Standard-Grammatik der englischen Sprache folgten (z. B. Imperative, Deskriptive, Interrogative), jedoch auch Äußerungen, die nur ein einzelnes Wort oder eine Nominalphrase enthielten, wie sie in Antworten auf Fragen zu erwarten sind.

Hier zeigt sich, dass fast alle grammatikalisch korrekten Äußerungen auch als semantisch kohärent bewertet wurden. Dieses Ergebnis war so zu erwarten, denn in der Regel sind das die Äußerungen, die vom Spracherkenner auch korrekt erkannt werden konnten. In den Äußerungen, die vom Sprachverstehen mit *nicht verständlich* bewertet wurden, kamen in fast allen Fällen dem Sprachverstehen unbekannte Wörter vor („toy-cow“ oder „colored“), die daher nicht mit den anderen Wörtern zu einer einzigen Ableitung verlinkt werden konnten.

Die Eingaben an das Sprachverstehen, die der Standard-Grammatik nicht folgen, sind entweder aufgrund von spontansprachlichen Phänomenen grammatikalisch nicht korrekt oder wurden vom Spracherkenner (in Teilen) falsch erkannt. Hier werden wie zu erwarten, die Äußerungen in vielen Fällen als nicht kohärent bewertet. Dabei ist die Bewertung der Ergebnisse nicht so eindeutig zu treffen. Als nicht verständlich bewertet wurden die vom Spracherkenner erkannten Äußerungen, die keine klare Bedeutung abbildeten. Hier handelte es sich zumeist um Fehl-Erkennungen, die an das Sprachverstehen weitergeleitet wurden, wie z. B. die falsch erkannte Äußerung „you it this can“, die aus einem falsch erkannten Meta-Kommentar wie z. B. „glad to hear that“ erzeugt wurde.

Die Äußerungen, die nur teilweise verstanden wurden, sind nicht so eindeutig zuzuordnen. Hier wurden die Äußerungen vielfach zum Teil falsch erkannt. Typische Beispiele für die erkannten Äußerungen sind „this stop“ anstatt „please stop“ oder „hello cube“ anstatt „yellow cube“. In geringerem Umfang konnten die Äußerungen auch nur deshalb teilweise analysiert werden, weil einzelne Wörter im Lexikon fehlten, die vom Spracherkenner jedoch korrekt erkannt werden konnten. Hier liefern die Analyseergebnisse lediglich einen Hinweis, dass das Ergebnis keine zuverlässige Interpretation darstellt. Ob und welcher Teil der Äußerung korrekt verstanden (und auch erkannt) werden konnte, kann nicht alleine aufgrund der Sprachanalyse getroffen werden. Hier muss Kontextinformation einbezogen werden.

Äußerungen, die grammatikalisch nicht korrekt aber als semantisch kohärent bewertet wurden, weisen verschiedene Ursachen auf. Hier handelt es sich zum Teil um Äußerungen mit einer ungewöhnlichen syntaktischen Struktur, wie sie oft in der *Baby-Sprache* oder im *Foreigner-Talk*

verwendet werden (z. B. „look cow“). Häufig sind dies auch Äußerungen, bei denen ein Wort nicht zum Rest der Äußerung passt. In diesen Fällen kann man davon ausgehen, dass eine Fehl-Erkennung vorlag, wie in der falsch erkannten Äußerung „this is a cube let“. Dabei werden entweder Teile der Äußerung fälschlicherweise als Störgeräusche missinterpretiert oder umgekehrt die Umgebungsgeräusche als Bestandteil der Äußerung. Auch werden ggf. einzelne Wörter aufgrund von Störfaktoren falsch erkannt. Genauso werden zu Beginn einer Äußerung häufig Wörter abgeschnitten, so dass nur Teile erkannt werden (wie in „is a cube“). Vielfach werden auch Namen von Objekten falsch erkannt. Dann entstehen Ergebnisse aus der Spracherkennung wie „cube let“. Hier ist es allein aufgrund der Sprachdaten nicht möglich genau zu erkennen, ob tatsächlich ein „Würfel“ (engl. „cube“) gemeint ist, oder ein anderes Objekt, das fälschlicherweise aus dem Signal erkannt wurde, aus dem „cube let“ erzeugt wurde. Hier wird deutlich, wie wichtig die Integration der visuellen Informationen aus der Szene ist. Die genaue Zuordnung der Äußerungen zur Bewertung ist in Tabelle 10.3 dargestellt. Dabei wird einerseits dargestellt, wieviele Äußerungen wie weit analysiert werden konnten und andererseits wie diese Ergebnisse in Beziehung zur Grammatikalität der interpretierten Einheiten stehen.

erkannte Äußerungen (1642)	gramm. korrekt (775 ≙ 47,2%)	ungramm. (867 ≙ 52,8%)	insgesamt (100%) (1642)
nicht korrekt [0,0–0,5[	5 (0,3%)	141 (8,6%)	146 (8,9%)
partiell korrekt [0,5–0,75[	25 (1,5%)	336 (20,5%)	361 (22%)
korrekt [0,75-1,0[	16 (1,0%)	98 (5,9%)	114 (6,9%)
vollständig korrekt [1,0]	729 (44,4%)	292 (17,8%)	1021 (62,2%)

Tabelle 10.3.: Ergebnisse der semantischen Verarbeitung nach interner Bewertung und syntaktischer Struktur aufgeschlüsselt

Vielfach handelt es sich bei den erkannten Äußerungen um einzelne Wörter, oftmals bestehen sie nur aus einem Verb oder Nomen wie z. B. „look“ oder „cup“. Hier liefert die semantische Kohärenz keine Hinweise, ob die Äußerung richtig erkannt wurde oder nicht. Nur mit Hilfe der Kontextinformation kann eine sichere Entscheidung getroffen werden. Zusätzlich bietet der zugehörige Dialogakt eine weitere Entscheidungshilfe, ob die Äußerung einen Sinn ergibt. Dann kann der Dialogmanager unter Einbeziehung der Situation entsprechend reagieren. Statt „hello“ wurde regelmäßig „yellow“ erkannt. Für „yellow“ existiert jedoch im Gegensatz zu „Hello“ kein Dialogakt, dieser wird dann als *fragment* angegeben, der Dialogmanager startet daher in diesem Fall eine Rückfrage. In 13% der Fälle (5% grammatikalisch korrekt, 8% ungrammatikalisch) konnte kein Dialogakt für die interpretierte Äußerung angegeben werden. Diese Äußerungen konnten daher nicht vom Dialogmanager verarbeitet werden, sondern eine Rückfrage war notwendig. Dies betraf sowohl Äußerungen, die eine schlechte Bewertung vom Sprachverstehen bekamen, als auch solche, die als semantisch kohärent eingestuft wurden (in 6% der Fälle, zumeist Einwort-Äußerungen).

Das Sprachverstehen unterstützte zusätzlich das gesamte System durch Hinweise auf Gesten. Diese wurden in allen Fällen korrekt weitergegeben und die Verbindung zwischen Sprache und

Bildinformationen konnte dadurch hergestellt werden. Aufgrund der recht langen Berechnungszeit (mehrere Sekunden) für die Erkennung von Objekten wurde die Kommunikation in vielen Fällen verzögert [Haa05]. Ohne die Hilfen aus dem Sprachverstehen hätten sich die Verarbeitungszeiten weiter verlängert, so konnte zumindest das Zeitfenster für den Erkennungsprozess des Objektes stark eingeschränkt werden.

10.3.4. Fazit

Die Evaluation des Experimentes mit dem im Robotersystem BIRON intergrierten Sprachverstehen belegt den sinnvollen Einsatz in Mensch-Roboter-Kontexten. Das Sprachverstehen liefert in Echtzeit für den Dialog sinnvolle Interpretationen der Äußerungen, insbesondere unter den besonderen Rahmenbedingungen der Mensch-Roboter-Kommunikation. Das Sprachverstehen liefert neben semantischen und pragmatischen Informationen zusätzlich eine Hilfestellung zur Bewertung der Spracherkennungsergebnisse. Darüber hinaus konnten viele Äußerungen, die von der Spracherkennung nur zum Teil korrekt erkannt wurden, dennoch vom Sprachverstehen in eine sinnvolle semantische Interpretation überführt werden. Ebenso konnten auch ungrammatikalische Äußerungen, die aufgrund der freien Kommunikationssituation auftraten, sinnvoll interpretiert werden. Beispielsweise wurden Äußerungen wie „look here BIRON look left“ (interpretiert als „Looking: (Potental_gesture: here, Person: BIRON, Orientation: left)“), „a book this is“ (interpretiert als „Existence: (Object: book, Maybe_gesture: this)“) oder „look cup“ (interpretiert als „Looking: (Object:cup)“).

Der begrenzende Faktor für die gesamte Kommunikationssituation ist vor allem die Spracherkennung, die mit einen kleineren Sprachumfang arbeitet als das Sprachverstehen. Die häufigen Fehler der Spracherkennung verhinderten, dass eine wirklich freie Interaktion entstehen konnte ([Kle04], S. 118 ff.). Es zeigt sich jedoch auch, dass für die Probanden gerade die sprachlichen Fähigkeiten ein wichtiges Anliegen an das Robotersystem darstellen [Wre04a, Li04]. Die Weiterentwicklung sprachlicher Fähigkeiten sind demnach eine wesentliche Aufgabe für die Konstruktion eines *Robot Companions*. Dies betrifft alle an der Kommunikation beteiligten Komponenten. Die Spracherkennung nimmt jedoch eine besondere Stellung ein, sie bildet das „Nadelöhr“ – schlägt die Erkennung fehl, können auch die anderen beteiligten Prozesse nicht wirken. Für die gesamte Interaktion läßt sich zukünftig hoffen, dass sich die Spracherkennung ggf. durch mehrere mit verschiedenen Wissensbasen ausgestatteten (getriggert durch den Themendetektor) parallel verwendeten Spracherkennern deutlich verbessern läßt und somit das Sprachverstehen auch im realen Robotersystem bis an seine Grenzen ausgereizt werden kann.

10.4. Vergleich mit syntaktischem Parsing

In diesem Abschnitt wird der in dieser Arbeit vorgestellte Ansatz mit klassischen Parsing-Strategien verglichen, genauer, die Performanz im Vergleich zur syntaktischen Analyse [Hüw06a]. Hierfür wurden wiederum die Daten aus dem Experiment mit dem Robotersystem BIRON verwendet, die im vorhergehenden Abschnitt bereits im Gesamtzusammenhang mit allen im Kommunikations-Prozess beteiligten Komponenten evaluiert wurden. Im Gegensatz zum vorherigen Abschnitt wurden direkt die von der Aufmerksamkeitssteuerung des Robotersystems empfangenen Äußerungen transkribiert und als Eingabe für das Sprachverstehen verwendet. Die Spracherkennung wurde innerhalb des gesamten Prozesses durch manuelle Eingaben simuliert, um so eine Erkennungsrate von 100% nachzubilden. Von den 1642 spontansprachlichen Äußerungen wurden zufällig 418 Äußerungen für die manuelle Transkription und syntaktische Analyse ausgewählt. Dabei wurde in 51 der Äußerungen festgestellt, dass ein entsprechender Lexikoneintrag in der verwendeten Datenbank für mindestens ein Wort fehlt.

Bei einer Rate von 48% korrekt erkannten Äußerungen (vgl. Kap. 10.3) konnte das Sprachverstehen in 61% der Fälle die Äußerungen vollständig semantisch interpretieren. Bei einer Erkennungsrate von 100% der spontansprachlichen Äußerungen lag die Rate immerhin bei 84% (Fehler verursacht aufgrund unbekannter Wörter und Meta-Kommentare). In Tabelle 10.4 sind die Ergebnisse im Vergleich dargestellt.

Sprachverstehen	Spracherkennung = 100%	Spracherkennung = 48%
nicht / teilweise interpretierbar	16%	39%
vollständig interpretierbar	84%	61%

Tabelle 10.4.: Interpretationsrate des Sprachverstehens bei einer vollständigen Erkennung der Äußerungen der Spracherkennung und verglichen mit der Erkennungsleistung aus den Experimenten aus dem Experimentsetting in Abschnitt 10.3.

Dieses Ergebnis wurde mit einem klassischen Parsing-Ansatz verglichen. Hierfür wurde jedoch kein realer Parser verwendet, sondern die syntaktische Korrektheit bewertet. Wie auch im Abschnitt zuvor, wurden Äußerungen als syntaktisch korrekt bewertet, die der englischen Syntax oder einer gängigen Antwort in Form einer Nominalphrase oder Einwort-Äußerung entsprachen („ja“, „hallo“, usw.). Um eine mögliche falsche Zuordnung zu verhindern, bewerteten drei Personen die Äußerungen anhand ihrer syntaktischen Korrektheit nach diesen genannten Kriterien. Dabei wurden von den 418 Äußerungen 314 als syntaktisch korrekt bewertet, von denen in 28 Fällen ein Lexikoneintrag fehlt, die also nicht vollständig syntaktisch erfasst werden konnten. Dagegen wurden 104 Äußerungen als nicht-korrekt klassifiziert. Davon wurden 58 der Äußerungen entweder vorne oder hinten abgeschnitten, was aufgrund der falschen Steuerung des aufgenommenen Eingabesignals durch die Aufmerksamkeitssteuerung des Robotersystems entstand. Beispiele sind hierfür Eingaben wie „can you find“, „is a cube“. Ebenso enthält es Instanzen, bei denen die Probanden sich selbst unterbrachen.

Äußerung	SSU	Dialogakt	syntaktisch korrekt
can you show me the book	Showing	instruction	+
can you see the cup	Looking	instruction	+
look biron this	Looking	instruction	-
look this	Looking	instruction	-
biron look	Looking	instruction	+
do you see	Looking	instruction	-
and look here	Looking	instruction	-
biron look here a	Looking	instruction	-
can you follow me to *the*	Move_afterwards	instruction	-
let us go	Move_to	instruction	+
give me a cup of coffee please	Giving	instruction	+
is a cube	Existence	description	-
here this is a cup	Existence	description	+
yes there is a cup	Existence	description	-
what can you do biron	Question_action	query	+
ho what can you do	Question_action	query	-
what can you	Question_action	query	-
what **choice** do i have	Question_action	query	+
so where is the cube	Question_position	query	+
how are you	Greeting	socialisation	+
i said hi biron	Greeting	socialisation	+
biron	Name_personal	fragment	+
this	Maybe_gesture	fragment	+
funny	Attribute	fragment	-
Legende: **unbekanntes Wort** *unverlinktes Wort*			

Tabelle 10.5.: Die Auswertung der Spontansprache aus den Ergebnissen der Spracherkennung – vollständig und korrekt interpretierte Äußerungen

In 68% der Fälle waren die Äußerungen syntaktisch korrekt und hätten daher aller Wahrscheinlichkeit nach mit einem klassischen Parsing-Mechanismus verarbeitet werden können. [4] Im Gegensatz dazu konnte das Sprachverstehen in 84% der Fälle eine sinnvolle semantische Interpretation gewinnen. In Tabelle 10.5 sind einige korrekt interpretierte Äußerungen abgebildet, zusammen mit der SSU, die die Wurzel darstellt, sowie dem zugehörigen Dialogakt und der syntaktischen Einschätzung. Ebenso wurden einzelne Wörter korrekt einer SSU zugeordnet, die sich jedoch nicht einem Dialogakt zuordnen ließen. 66 Äußerungen konnten nur teilweise (siehe Tabelle 10.6) oder gar nicht interpretiert werden (siehe Tabelle 10.7). Zum Zeitpunkt der Evaluation wurden die Äußerungen noch mit der einfachen links-rechts Strategie verarbeitet. Mit der

[4] Komplexere Mechanismen die auch einige spontansprachliche Äußerungen verarbeiten können, wie z. B. in [Kro00] oder [McK98] beschrieben, würden wahrscheinlich bessere Ergebnisse erzielen, benötigten jedoch entsprechende Kenntnisse über die möglichen Äußerungen.

Proximal-Suche würde aus der Äußerung „glad to hear that hi biron what can you do“[5] mehrere Teil-Ableitungen erzeugt werden, wobei nur die erste aufgrund fehlender Lexikoneinträge eine falsche Interpretation darstellt (siehe Abb. 10.13).

Äußerung	SSU	Dialogakt	synt. korrekt
should i take **anything else off**	Take	instruction	+
is there **anything else** you *can do*	Action	instruction	+
let us **start** *over* look here	Looking	instruction	-
it is **very nice**	Existence	description	+
ey the cow is **very nice**	Existence	description	+
this is a **p**	Existence	description	-
am i am *doing* **anything else**	Existence	description	+
we are **getting somewhere**	Existence	description	+
what is the **wrong**	Question_action	query	-
glad *to* **hear** *that* hi biron *what can* you *do*	Greeting	socialisation	-
can you **identify** the keyboard	Object_office	object	+
can you **find** the cube	Object	object	+
not **yet**	Negation	fragment	+
no *you not*	Negation	fragment	-
that i do not **understand** you	Negation	fragment	-
and *this*	Adding	fragment	-
ah **actually** *biron*	Exclamation	fragment	-
biron you **still** there	Position	fragment	-
you **understand** *me*	Proxy_personal	fragment	-
Legende: **unbekanntes Wort** *unverlinktes Wort*			

Tabelle 10.6.: Die Auswertung der Spontansprache aus den Ergebnissen der Spracherkennung – teilweise interpretierte Äußerungen

In 20 Fällen führten fehlende Lexikoneinträge dazu, dass Äußerungen nur teilweise interpretiert werden konnten und in 8 Fällen zu keiner semantischen Interpretation. In 23 Fällen, in denen ein Wort unbekannt war, konnte das System dennoch eine vollständige semantische Zuordnung liefern. Die Äußerung „can you go for a walk with me“ (deutsch: „gehst du mit mir spazieren“) wurde interpretiert als „can you go with me“, wobei „for a walk“ mit dem unbekannten Wort „walk“ nicht integriert wurde. Ebenso wurde die Äußerung „can you come closer“ interpretiert als „can you come“, wobei das unbekannte Wort „closer“ nicht berücksichtigt wurde.

Weitere unbekannte Wörter sind:

> glad, should, off, else, yet, actually, anything, getting, still, identify, understand, very, nice, wrong, great, so, usual, good, sound, somewhere ...

In 21 Fällen (5%) enthielten die Interpretationen falsche semantische Zuordnungen oder der Dialogakt wurde nicht korrekt zugeordnet (3%) (siehe Tabelle 10.8). Das Wort „okay“ oder „not“

[5] entstanden aufgrund fehlgeschlagener Segmentierung der Äußerung

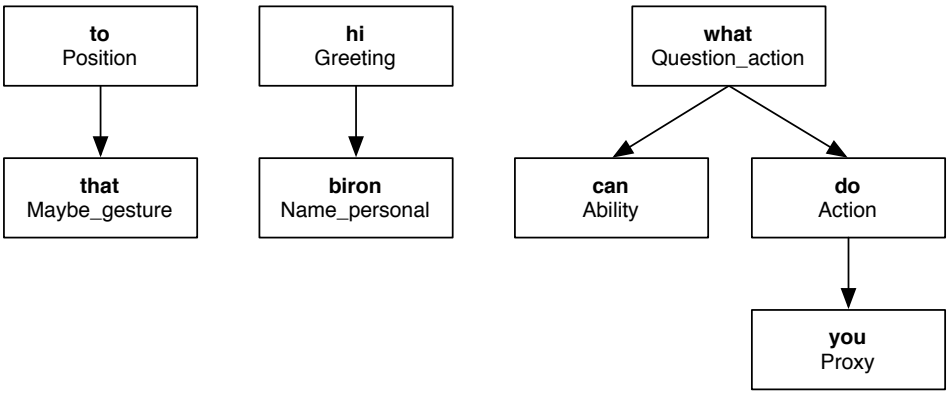


Abbildung 10.13.: Ableitungstrukturen der Äußerung „glad to hear that hi biron what can you do“

Äußerung	SSU	Dialogakt	synt. korrekt
ain‘t usual *but* **still**	UNKNOWN	fragment	-
ain’t *is*	UNKNOWN	fragment	-
sound good *this*	UNKNOWN	fragment	-
sorry	UNKNOWN	fragment	+
Legende: **unbekanntes Wort** *unverlinktes Wort*			

Tabelle 10.7.: Die Auswertung der Spontansprache aus den Ergebnissen der Spracherkennung – nicht erkannte Äußerungen

wurde beispielsweise fälschlicherweise gar keinem Dialogakt zugeordnet statt des Dialogaktes *Confirmation* oder *Negation*. Die Äußerung „at is fine“ mit dem unbekannten Wort „fine“ wurde interpretiert als „is at“, also eine Teil-Beschreibung einer Position, bei der das Objekt fehlt. Des Weiteren wurde „here we are getting somewhere“ mit der unbekannten Phrase „getting somewhere“ fälschlicherweise interpretiert als „we are here“, ebenso die Äußerung „hey we are done here“. Die nächste Äußerung „too difficult to tell let us try that one biron look here a cube“ entstammt wiederum mehreren nicht segmentierten linguistischen Einheiten. Ihre Interpretation entspricht in etwa der Bedeutung der Aussage „look at us biron – this here – a cube“. Aus der Aussage „please do not tell me it is my fault“ wurde eine Interpretation generiert, die in etwa der Aussage „please do not tell me my (name) is (x)“ entspricht. Mit der Proximal-Suche entspricht es der Ableitung in Abbildung 10.14. Die beiden letzteren Äußerungen ergeben keinen Sinn. Das Wort „like“ wurde nicht identifiziert mit „mögen“, da dieser Verweis im Lexikon nicht existiert, sondern mit „wie“.

Unter der Annahme, dass eine syntaktisch korrekte Äußerung auch zu einer semantisch korrekten Interpretation führt, kann die syntaktische Korrektheit mit der semantischen Analyse verglichen werden. Wie in der Tabelle 10.9 zu sehen, konnte in 84% der Fälle unter Einsatz des in dieser

Äußerung	SSU	Dialogakt	synt.
at is **fine**	Existence	description	-
here we are **getting somewhere**	Existence	description	+
hey we are *done* here	Existence	description	+
too difficult to *tell* let us **try** that one biron look here a cube	Looking	instruction	-
please do not tell me it is my **fault**	Negation	fragment	+
do not **be embarrassed** with me	Negation	fragment	+
do you like *it*	Conjunction_modal	fragment	+
Legende: **unbekanntes Wort** *unverlinktes Wort*			

Tabelle 10.8.: Die Auswertung der Spontansprache aus den Ergebnissen der Spracherkennung – falsch erkannte Äußerungen oder falsche Zuordnung des Dialogaktes

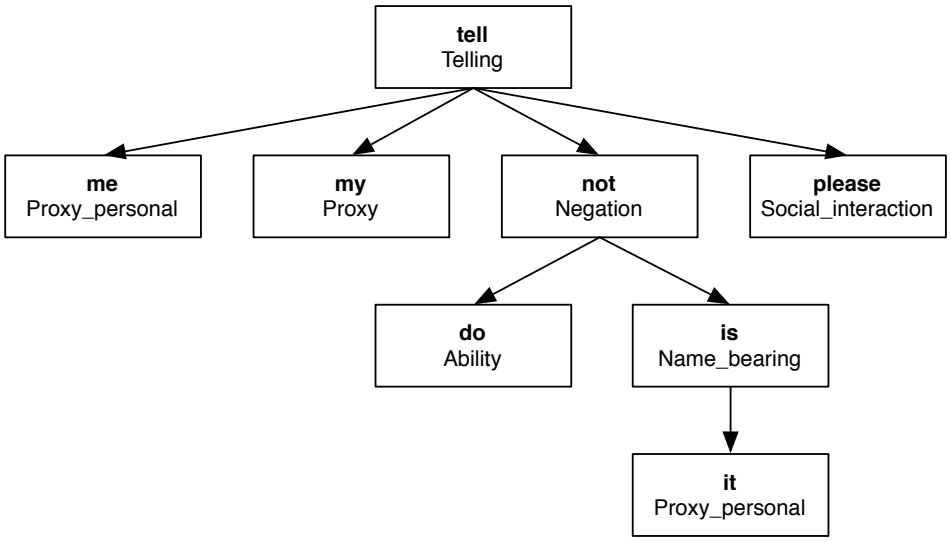


Abbildung 10.14.: Ableitungstrukturen der Äußerung „please do not tell me it is my fault“

Arbeit beschriebenen Verarbeitungsmechanismus eine vollständige semantische Interpretation gewonnen werden und nur in 16% der Fälle eine partielle oder gar keine. Dagegen entsprachen nur 68% der Fälle den syntaktischen Regeln, 32% der Äußerungen folgten dagegen den syntaktischen Vorgaben nicht. Dies lässt vermuten, dass die semantische Verarbeitung ungefähr die Hälfte der Äußerungen zusätzlich abdeckt, die mit einer syntaktischen Analyse nicht verarbeitet werden können. Aus den Ergebnissen läßt sich folgern, dass gerade dann, wenn nicht vollständig vorherzusagen ist, wie die Benutzer sich äußern, sie jedoch mit großer Wahrscheinlichkeit unbekannte Wörter verwenden und sich spontansprachlich ausdrücken, dieser Ansatz des Sprachvertehens geeignet ist, um semantische Interpretationen aus den Äußerungen der Interaktionspartner eines mobilen Robotersystems zu gewinnen.

	Sprachverstehen		synt. Korrektheit
nicht / teilweise interpretierbar	16%	nicht korrekt	32%
vollständig interpretierbar	84%	korrekt	68%

Tabelle 10.9.: Vergleich der semantischen Verarbeitungsergebnisse mit der syntaktischen Korrektheit, unter Annahme einer 100% Spracherkennungsrate

10.5. Zusammenfassung

In diesem Kapitel wurde das Konzept des Verstehens von situierter Spontansprache unter Berücksichtigung verschiedener Aspekte evaluiert. Die Evaluation einer einzelnen Komponente, integriert in einem Gesamtsystem, ist nicht ganz unproblematisch, da die einzelnen Leistungen der Komponente eng mit der gesamten Funktionalität des Systems verwoben sind. Bei der Interaktion mit einem Robotersystem fließen immer die Fähigkeiten und auch die Grenzen der anderen am (Mensch-Roboter) Dialog beteiligten Komponenten mit ein. Daher wurde die Evaluation des Sprachverstehens in mehrere Unterbereiche aufgeteilt. Die Evaluationsergebnisse der einzelnen Evaluationen sind Tabelle 10.10 kurz zusammengefasst.

Die Untersuchungen haben die Leistungsfähigkeit des Ansatzes unterstrichen. Es hat sich gezeigt, dass das System gut geeignet ist, Äußerungen in dem Kontext von Mensch-Roboter-Kommunikation zu analysieren und geeignete semantische (und pragmatische) Interpretationen zu erzeugen. Es ist fähig, eine große Bandbreite an Äußerungen zu verarbeiten, sowohl komplexe Hauptsätze als auch situierte und spontane Äußerungen. Die Äußerungen der Probanden aus dem Hometour-Korpus konnten weitestgehend interpretiert werden, nur in wenigen Fällen stößt das System hier an seine Grenzen. Der sinnvolle Einsatz des Sprachverstehens im Gesamtsystem wurde gezeigt: gerade auch dann, wenn die Spracherkennungsleistung weniger gut ist. Der Dialogmanager reagierte aufgrund der Interpretationen der Äußerungen auch unter schwierigen Bedingungen sinnvoll. Die Evaluation der Experimente mit dem Robotersystem zeigten, dass auch im laufenden Betrieb das Sprachverstehen das Robotersystem substanziell unterstützen kann. Die Sprachverstehenskomponente kann für das Dialogsystem sinnvolle Interpretationen aus den Äußerungen gewinnen, auch wenn diese nicht vollständig korrekt erkannt werden oder spontansprachliche Phänomene enthalten. Ebenso bewertet die Verstehenskomponente die Ergebnisse des Spracherkenners sinnvoll und erhöht somit die Robustheit des gesamtem Systems.

Insgesamt zeigt sich, dass die in dieser Arbeit vorgestellten Datenstrukturen und Verarbeitungsmechanismen einen geeigneten Ansatz für das automatische Verstehen von Spontansprache im Kontext der *Robot Companions* darstellt. Das vorgestellte Verfahren ist effizient und hat sich sowohl in der Praxis als auch in synthetischen Experimenten bewährt.

Merkmale / Funktion	Daten	Online / Offline	Qualitative / Quantitative Analyse	Fazit / Ergebnis
Fähigkeiten allgemein (10.1)	artifiziell und natürlich	offline	qualitativ	große Bandbreite an Äußerungen
Verarbeitung von Spontansprache (10.2)	natürlich	offline	qualitativ	kann Spontansprache verarbeiten
Gesamtfunktionalität im Robotersystem (10.3)	natürlich	online	quantitativ	effizient, bewertet Güte der Spracherkennung, kann fehlerhafte Eingaben der Spracherkennung verarbeiten
Vergleich mit klassischen Ansätzen (10.4)	natürlich	online	qualitativ und quantitativ	bessere Ergebnisse als klassisches Parsing, verarbeitet auch Äußerungen mit unbekannten Wörtern, verarbeitet ungrammatikalische Äußerungen

Tabelle 10.10.: Die Ergebnisse der einzelnen Evaluationen im Überblick

11. Diskussion und Ausblick

Zunächst folgt in diesem Kapitel eine Auseinandersetzung mit dem hier vorgestellten Ansatz im Vergleich zu sonst üblichen syntaktischen Verfahren. Anschließend folgt der Ausblick auf mögliche Erweiterungen des hier vorgestellten Ansatzes zur Verarbeitung situierter Spontansprache im Mensch-Roboter-Kontext. Sie berücksichtigen auch in der Diskussion angeregte Erweiterungen.

11.1. Diskussion

Auch wenn in dieser Arbeit einzig die semantische Analyse für das Verstehen gesprochener Sprache verwendet wurde, bedeutet das nicht, dass die bisher bekannten grammatikalischen Ansätze zur Verarbeitung von Sprache unterbewertet werden sollen. Sie sind für verschiedene Problemstellungen von großer Bedeutung. Der hier vorgestellte Ansatz soll vielmehr die Anregung zu neuen Methoden und Konzepten liefern. Es soll argumentiert werden, dass es wichtig ist, sich in bestimmten Kontexten vom rein grammatikalischen Ansatz zu lösen. Gerade die gesprochene Sprache ist nicht immer formal und korrekt. Daher bietet der hier gezeigte Ansatz einen neuen Weg - anstatt die Äußerungen mittels einer formal definierten Grammatik zu analysieren - über die semantischen und pragmatischen Zusammenhänge zu gehen. Letztendlich können sowohl Syntax als auch Semantik bei der Verarbeitung ihren Anteil beitragen und Wissen über die mögliche Interpretation bereitstellen. In vielen Fällen ist es gut möglich, syntaktische Strategien bei der Verarbeitung zu nutzen und darauf aufbauend semantische Informationen zu gewinnen. Jedoch stößt dieser Ansatz vor allem bei Spontansprache an seine Grenzen, da dort andere Regeln als bei geschriebenen Texten gelten, die syntaktischen Regeln sozusagen aufgeweicht sind. Gerade wenn das Ziel die Gewinnung von semantischen Informationen ist, wie bei sprachlichen Mensch-Maschine-Schnittstellen, so sind syntaktische Analysemethoden nur ein Vehikel, um an die semantischen Interpretationen zu gelangen. Wie in dieser Arbeit gezeigt wurde, kann daher die umgekehrte Vorgehensweise in bestimmten Kontexten sinnvoller sein: von dem semantischen Wissen auszugehen und darauf aufbauend die Äußerungen zu analysieren (vgl. insbesondere Kap. 10.4). An Stellen, an denen die Verarbeitung an ihre Grenzen stößt, können dann jeweils zusätzlich syntaktische Informationen genutzt werden, um komplexere Sprachphänomene aufzulösen (z. B. bei der Anaphernresolution). Um dennoch die notwendige Robustheit zu gewährleisten, kann dabei der Schwerpunkt auf einzelnen Phrasen liegen, anstatt die gesamte Äußerung auf syntaktische Korrektheit zu analysieren. Im nächsten Abschnitt wird diese Idee genauer ausgeführt. Jedoch lohnt sich zur Zeit die Integration zusätzlicher syntaktischer Informationen nicht, da, wie in Kapitel 10 gezeigt, die rein semantische Analyse für den Korpus der Mensch-Roboter-Interaktion vollkommen ausreicht.

Aus der Konzeption der *situierten semantischen Einheiten* wurde ersichtlich, dass Syntax und Semantik in enger Beziehung zueinander stehen. In der semantischen Zuordnung einzelner Wörter spiegeln sich ebenfalls syntaktische Eigenschaften wider. In die semantische Verarbeitung können demnach auch syntaktische Informationen einfließen, wie in Abschnitt 11.2 diskutiert wird. Die beiden Ebenen der Syntax und Semantik enthalten demnach parallel in Teilen dieselben Informationen. Beispielsweise stellt der syntaktische Unterschied zwischen einem Personalpronomen und einem Possessivpronomen („ich“ vs. „mein“) semantisch auch unterschiedliche Informationen dar. Das Possesivpronomen enthält neben der Information, dass es sich um eine Person handelt, zusätzlich die Information einer besitzenden Person. Das heißt, in diesem Fall werden diese unterschiedlichen Informationen ebenfalls durch unterschiedliche SSUs dargestellt. In anderen Fällen scheint es, dass semantische Unterschiede nicht durch syntaktische Unterschiede zu erkennen sind und umgekehrt. Wie genau die semantischen Feinheiten dargestellt werden sollten, hängt jedoch auch von der Anwendung und dem Kontext ab. Beispielsweise ist der Genus oder Kasus von Nomen gerade in der Spontansprache und auch aufgrund von Spracherkennungsfehlern nicht immer genau zuzuordnen. Hier müssen (wie auch in [Kro00] beschrieben) Mechanismen gefunden werden, die in diesem Bereich Offenheit zulassen. In unserem System wird zur Zeit nicht auf den Genus geachtet, jedoch ist künftig denkbar, den Genus mit einfließen zu lassen. Dabei sollte zuerst auf den korrekten Genus geachtet, und nur wenn keiner vorhanden ist, andere syntaktisch nicht korrekte Genera im Verarbeitungsprozess berücksichtigt werden (ggf. durch eine schlechtere Bewertung). Dasselbe gilt für die Einbindung des Kasus. Aufgrund spontansprachlicher Äußerungen und möglicher Erkennungsfehler des Spracherkenners war die genaue syntaktische Unterscheidung in der hier vorliegenden Umsetzung des Konzeptes nicht sinnvoll. Wie genau diese syntaktisch-semantischen Beziehungen aussehen, sollte in der hier vorgestellten Arbeit jedoch nicht Gegenstand der Untersuchung sein.

Der Mechanismus ist nicht für alle Problemklassen gleich gut geeignet, wie in Abschnitt 10.1 gezeigt wurde. Dafür sind die jeweiligen Äußerungen und Sätze, die die verschiedenen Aufgaben mit sich bringen, zu unterschiedlich. Bei der Analyse wohlgeformter Sätze mit komplexer Struktur, wie z. B. im Bereich der Analyse von Zeitungsartikeln oder ganzen Romanen, sind syntaktisch getriebene Mechanismen vermutlich besser geeignet, da sie eher die grammatikalische Struktur und damit feinere semantische Bedeutungen abbilden können. Umgekehrt jedoch lassen sich spontansprachliche Äußerungen in einem recht freien Kontext nicht gut mit rein syntaktischen Methoden analysieren. Die Unterschiede zwischen wohlgeformten Sätzen und freier Sprache ist sehr groß, was bei dem Versuch, Spontansprache mit konventionellen Ansätzen zu analysieren, sehr deutlich wird. Hier sind komplett neue Wege gefragt. Diese Arbeit soll genau hierzu einen Beitrag leisten und zum einen für die Problematik der Mensch-Roboter-Kommunikation sensibilisieren und gleichzeitig eine Möglichkeit zur Lösung der besonderen spontansprachlichen Anforderungen liefern. Die Tabelle 11.1 zeigt noch einmal die verschiedenen Systeme aus Kapitel 4 im Vergleich mit dem in dieser Arbeit vorgestellten Ansatz der Verarbeitung von situierter Spontansprache. Im Vergleich zu den anderen Systemen wird deutlich, dass dieser Ansatz die gesamte Problematik der Mensch-Roboter-Kommunikation weitestgehend berücksichtigt und besonders geeignet ist, diese Art der Dialoge zu verarbeiten. Jedoch lassen sich Erweiterungen in das Gesamtkonzept integrieren, die das Sprachverstehen zwar komplexer, jedoch auch mächtiger machen können. Eine Auswahl an vielversprechenden Erweiterungen werden im Folgenden diskutiert.

System	Eigenschaften					
	Mechanismus	**Sprachumfang**	**Spracheingabe**	**Spontansprache**	**Situierte Sprache**	**Echtzeitfähigkeit**
SHRDLU	synt. Parser	wenige reg. Ausdrücke	Tastatur	−	−	−
Verbmobil	komplexe Analyse	?	Mikrofon	+	−	?
CORA	komplexe Analyse	Handlungsanweisungen	Tastatur	+	(+)	+
TJ	Mustersuche	wenige reg. Ausdrücke	Tastatur	−	−	?
MOBSY	Mustersuche	reg. Ausdrücke	Mikrofon	−	−	+
JIJO-2	synt. Parser	200 Wörter, 90 Regeln	Mikrofon	−	−	+
Projekt IBL	synt. Parsing	ca. 200 Wörter	Mikrofon (indirekt)	−	+/−	?
CARL	LFG-Parser	36 Wörter	Mikrofon	−	−	+
Flo	Keywordspotting	100 Wörter	Mikrofon	−	−	+
System in [Bau01]	synt. Parser + sem. Netze	Handlungsanweisungen	Mikrofon	+	+	+
KAMRO	synt. Parser	?	Mikrofon	−	(+)	+
Hermes	komplexe Analyse	Kommandosätze	Tastatur o. Mikrofon	−	(+)	+
WITAS	synt. Parser	70 Wörter	Graf. Interface u. Mikrofon	−	(+)	+
BIRON	sem. Analyse	1400 Wörter	Mikrofon	+	+	+

Tabelle 11.1.: Vergleich der Systeme nach ausgewählten Kriterien

11.2. Ausblick

Der in dieser Arbeit vorgestellte Ansatz zum automatischen Verstehen gesprochener Sprache wurde von Beginn an mit besonderer Rücksicht auf Modularität und Erweiterbarkeit entwickelt. So lassen sich weitere Mechanismen im Basiskonzept integrieren, die zu einer noch besseren Interpretation von Spontansprache führen können. Dabei muss jedoch ebenfalls berücksichtigt werden, dass der Prozess des Sprachverstehens nicht für sich allein, sondern immer im Gesamtsystem betrachtet werden sollte. So können Defizite der Spracherkennung bei einer Integration von komplexeren Mechanismen, insbesondere zur Verarbeitung von Ellipsen und Anaphern, durchaus größere Probleme hervorrufen als in der Basisversion, die geringere Ansprüche an die Leistung des Spracherkenners stellt. Durch diesen Unsicherheitsfaktor kann das Gesamtsystem sogar an Robustheit verlieren. Es ist also im Vorfeld abzuschätzen, wie zuverlässig die Ergebnisse aus der Spracherkennung sind. Mitunter ist es hilfreich innerhalb einer umfassenden Evaluation der Spracherkennung festzustellen, welche Arten von Fehl-Erkennungen mit welcher Häufigkeit auftreten. Danach kann abgeschätzt werden, ob das Einbinden weiterer Methoden eine Verbesserung des gesamten Dialogsystems zur Folge hat.

Die Restunsicherheit beim Sprachverstehen lässt sich vermutlich nicht vollständig beseitigen. Daher ist es nicht zu verhindern, dass das Robotersystem bei unsicherem Sprachverständnis Rückfragen stellen muss, um eine Fehl-Handlung zu vermeiden. Dennoch sollte die Anzahl der Fragen möglichst gering gehalten werden, um eine weitgehend reibungslose Kommunikation zu erlauben. Für dieses Ziel werden im Folgenden Vorschläge zur Erweiterung des Verstehensprozesses erörtert. Aufgrund der Ergebnisse lassen sich dabei drei Problembereiche festmachen.

Übergeben mehrerer Parsebäume

Gegenwärtig wird jeweils immer nur ein Parsebaum mit dem zugehörigen Dialogakt an den Dialogmanager weitergereicht. Aus den Erfahrungen zeigt sich jedoch, dass es durchaus sinnvoll sein kann, mehrere Parsebäume weiterzureichen. Mitunter werden von der Spracherkennung gleich mehrere Äußerungen an das Sprachverstehen geschickt. Diese werden aufgrund fehlender Separierungsmarkierungen wie eine einzige Äußerung behandelt. Zu erkennen sind diese „Mehrfachäußerungen“ daran, dass Teilbäume erzeugt werden, bei denen sich die zugehörigen Wortketten klar voneinander abgrenzen und sich nicht überschneiden (siehe Kap. 10.4). In diesen Fällen ist es sinnvoll alle Teilbäume an den Dialogmanager weiterzureichen, so dass alle Informationen verarbeitet werden können und nicht Teile der Äußerung verloren gehen.

Einbindung syntaktischer Informationen

Bei der Durchführung „synthetischer“ Studien ist aufgefallen, dass es Äußerungen gibt, deren Sinn sich mit rein semantischen Methoden nicht eindeutig erschließen lassen. Für die Analyse dieser Äußerungen sind zusätzliche Reihenfolgeinformationen wichtig. Auch wenn in den Be-

nutzerstudien mit dem Roboter BIRON solche Äußerungen nur selten aufgetreten sind, ist es erstrebenswert, diese auch korrekt analysieren zu können. In Kapitel 10.1.6 wurde dieses Problem anhand der Äußerung „biron follow sonja“ ausführlich erläutert. Dort wurde als Lösung für das Problem vorgeschlagen, Hinweise für die Suchstrategie des Verlinkungprozesses in die SSUs selbst zu integrieren. Dieses Konzept kann auf allgemeine, syntaktisch oder konzeptuell fundierte Beziehungen von Wörtern oder SSUs untereinander übertragen werden. Es kann sowohl auf der Ebene des Lexikons als auch der SSUs angewandt werden, wobei angesichts der Komplexität für konzeptuelle Beziehungen die zweite Variante definitiv vorgezogen werden sollte. Die Entkopplung von Lexikon und SSUs bietet hierbei eine enorme Flexibilität: Wenn eine Erweiterung einer SSU zu Widersprüchen bei einem Teil der ihr zugeordneten Lexikoneinträge führen sollte, kann diese SSU einfach aufgeteilt werden, um die Widersprüche aufzulösen. Auch wenn daraufhin Anpassungen an weiteren SSUs nötig werden, hält sich der Aufwand in einem überschaubaren Rahmen.

Auf der Ebene des Lexikons ist die Möglichkeit für eine syntaktische Plausibilitätsprüfung während oder nach der semantischen Analyse ohnehin bereits vorgesehen. Ergeben sich beim Parsevorgang Unsicherheiten, kann beispielweise die Zugehörigkeit eines Attributs zu einem Objekt durch eine Überprüfung der KNG-Kongruenz verifiziert werden. Dies kann insbesondere bei der Anaphernresolution von großem Nutzen sein. Zusammenfassend gilt: Syntaxbezogene Hinweise für den Verlinkungsprozess sollten in das Lexikon integriert werden und konzeptbezogene Hinweise in die SSU-Datenbank.

Implizites Wissen

Informationen, die ein Sprecher als allgemein bekannt voraussetzt, werden oft nicht mit geäußert (Konversationsmaxime nach Grice, siehe [Mei99]). Um dieses implizite Wissen in das Interpretations-Ergebnis zu integrieren, können bestimmte Erwartungswerte in den Verarbeitungsprozess eingebracht werden. Bei Imperativen fehlt beispielsweise der Akteur der Handlung, der implizit mit dem Interaktionspartner gleichgesetzt wird.

Im Roboterszenario ist der gesuchte Akteur in einem solchen Fall höchstwahrscheinlich der Roboter selbst. Dieser Erwartungswert kann als Vorbelegung oder „Default“ in dem Mechanismus mit einfließen. Das Verfahren kann man sich in etwa so vorstellen: Zuerst wird das Ergebnis der Interpretation auf Lücken überprüft. Fehlt einer SSU der zugehörige Verweis zur SSU *Actor*, so wird der Interpretationsprozess mit dem zusätzlich angefügten Wort „BIRON“ noch einmal wiederholt. Auf diese Weise können Imperative verarbeitet werden, ohne semantische Lücken im interpretierten Ergebnis zu hinterlassen. Für das Robotersystem BIRON ist dieses Verfahren allerdings nicht notwendig, weil der Dialogmanager diesen Spezialfall ohnehin selbst berücksichtigt.

Jedoch können mit dem hier vorgestellten Ansatz auch andere Informationen eingebunden werden, z. B. bereits bekanntes Wissen aus der Szene, Wissen aus der Dialog-Historie oder aus Antworten, die der Roboter selbst an die Benutzer richtet. Implizites Wissen dynamischer Natur

kann nicht mit Hilfe von „Defaults“ behandelt werden, sondern bedarf einer Infrastruktur, die sowohl auf den aktuellen Zustand als auch auf den bisherigen Verlauf des Diskurses mit allen beteiligten Modalitäten zugreifen kann.

Bei der Auflösung von Anaphern und Ellipsen befinden sich die notwendigen Informationen im einfachsten Fall innerhalb derselben Äußerung, die daher direkt innerhalb des Parseprozesses verarbeitet werden könnte. Eine Anapher oder Ellipse könnte sich aber auch auf eine vorherige Äußerung beziehen. Um auch für diesen Fall über die nötige Information zu verfügen, müsste der Parseprozess eine Liste der jeweils letzten Äußerungen verwalten. Dies ist leider immer noch nicht für jeden Fall ausreichend, da sich eine Anapher oder Ellipse nicht nur auf eine Äußerung des Benutzers sondern auch auf eine Rückfrage des Roboters beziehen kann. Also müssten auch die Rückfragen des Roboters in geeigneter Weise in die Liste integriert werden. Das alles berücksichtigt allerdings noch nicht, dass Informationen aus anderen Modalitäten beteiligt sein können, oder dass ein Themenwechsel innerhalb des Dialogverlaufes stattfinden kann (z. B. eine andere Person interagiert zwischenzeitlich mit dem Roboter). Um das Problem umfassend zu lösen, wird also nicht nur eine Vielzahl von Schnittstellen zu anderen Modulen der Roboter-Plattformen benötigt. Die verschiedenen Informationsquellen müssen auch entsprechend aufbereitet und verwaltet werden.

Innerhalb der BIRON-Plattform gibt es bereits ein Modul, das zwar nicht alle, aber schon einen Teil dieser Aufgaben übernimmt: Der Themendetektor fokussiert auf das aktuelle Thema und speichert und verwaltet die relevanten Informationen aus dem Dialogkontext. Die Komponente besitzt demnach Kenntnisse über die Aktionen und Objekte im aktuellen Geschehen sowie deren Zusammenhänge. Hierfür werden bereits einige der genannten Schnittstellen verwendet. Es ist also deutlich sinnvoller, den Themendetektor entsprechend zu erweitern, anstatt eine komplett neue Infrastruktur zu entwerfen und somit viel Funktionalität unnötig zu duplizieren. Hierfür müsste der Themendetektor nicht nur das aktuelle Thema, sondern auch den aktuellen Kontext auf unterschiedlichen Ebenen inklusive Historie bereitstellen. Der Parseprozess kann dann bei fehlender Information mit Hilfe einer Funktion `get_last_objects()` oder `get_last_actions()` versuchen, die Äußerung zu einer semantisch kohärenten Struktur zu vervollständigen.

12. Zusammenfassung

Thema der vorliegenden Arbeit ist die Konzeption und Entwicklung einer Sprachverstehenskomponente für die Mensch-Roboter-Interaktion.

Der Roboter dient in diesem Zusammenhang nicht als reiner Dienstleister, sondern ist eher als eine Art künstlicher Kompagnon mit sozialen Fähigkeiten zu verstehen. Dafür benötigt das Robotersystem sprachliche Fähigkeiten, die über die einfache Verarbeitung weniger Anweisungen weit hinausgehen. Der Roboter soll beispielsweise Fragen beantworten, Feststellungen interpretieren und Anweisungen entgegen nehmen können. Dazu gehört ebenfalls, dass sich seine sozialen Fähigkeiten in der Sprache widerspiegeln, z. B. soll er auf Begrüßungen und auf Höflichkeitsfloskeln angemessen reagieren. Er muss in der Lage sein, die Situation sowohl sprachlich als auch visuell zu erfassen und diese Informationen möglichst in sein bisheriges Wissen zu integrieren. Dazu benötigt das System ein umfangreiches Verständnis über die Sprache, mit der er konfrontiert wird: Er muss sowohl über eine interne Wissensrepräsentation seiner Domäne verfügen, als auch über die Möglichkeit, dieses Wissen zum Verstehen von Äußerungen und Dialogsituationen nutzen zu können.

Das Design der Sprachverstehenskomponente in dieser Arbeit entstand unter Berücksichtigung der vielfältigen Rahmenbedingungen speziell bezogen auf die Anforderungen eines solchen Roboterszenarios. Im Zentrum der Arbeit steht dabei die Entwicklung eines ganzheitlichen Ansatzes, bei dem nicht nur die technischen Aspekte und das Wissen über Robotersysteme, ihre möglichen Anwendungskontexte und die dafür eingesetzten Systemkomponenten eine Rolle spielen. Die verschiedensten Forschungsrichtungen – ausgehend von den soziologischen, psycholgischen und linguistischen Betrachtungen des Themas (vgl. Kap. 2) sowie die Einbeziehung des Wissens über Robotersysteme finden im gesamten Design Beachtung. Nur wenn möglichst viele Einflussfaktoren berücksichtigt werden, kann das Sprachverstehen in der Interaktion mit anderen Systemkomponenten eine intuitive Kommunikationssituation unterstützen.

Da die Entwicklung von Robotersystemen mit sozialen Fähigkeiten ein noch recht neues Forschungsgebiet darstellt, existieren nur wenige Studien, die sich mit den Besonderheiten der Kommunikation zwischen Mensch und Robotersystem befassen. Daher wurden in dieser Arbeit zunächst die generellen Kommunikationsaspekte betrachtet, andere Forschungsbereiche mit einbezogen und die dafür relevanten Aspekte auf diesen Kontext übertragen. Insbesondere die Erfahrungen aus dem Bereich der Mensch-Maschine-Kommunikation wurden in diesem Ansatz berücksichtigt. Um ein vollständiges Bild der Besonderheiten der Kommunikation zwischen Mensch und *Robot Companion* zu erhalten und die Unterschiede zu anderen Mensch-Maschine-Kontexten zu klären, wurden verschiedene Studien und Experimente zur Kommunikation zwi-

schen Mensch und Roboter erstellt und durchgeführt. Diese wurden insbesondere im Hinblick auf die sprachlichen Eigenheiten untersucht und bilden die Grundlage für die Wissensdatenbank des sprachverstehenden Systems.

Der Mensch-Roboter-Kontext setzt sich deutlich von anderen Kontexten, wie z. B. der telefonbasierten Interaktion oder der Interaktion über Tastatur, ab. Die Situiertheit und die Spontansprache spielen in diesen Szenarien eine besondere Rolle. Die Sprache bildet dabei ein breites Spektrum an sprachlichen Phänomenen und Dialogstilen ab und unterliegt im Gegensatz zu geschriebenen Texten nur selten festen grammatikalischen Regeln. Die Einbeziehung von Kontext- und Umweltinformationen sowie die Freiheit der Äußerungsformen, die die Interaktionspartner verwenden, prägen die Dialoge.

Ein weiterer zu berücksichtigender Bereich sind die direkten Schnittstellen zu der Sprachverstehenskomponente sowie auch die indirekten Schnittstellen und Informationsflüsse innerhalb der Roboterplattform. Hier ist insbesondere die Verbindung zur Spracherkennung zu nennen, da die Situationen, in denen sich mobile Roboter befinden, besondere Herausforderungen für die Erkennungsleistung darstellen. Fehl-Erkennungen sind unvermeidlich und diese Problematik muss auch im Designprozess des Sprachverstehens beachtet werden.

Aus den Besonderheiten der Mensch-Roboter-Kommunikation folgen die speziellen Anforderungen an das Konzept des Verstehens situierter Spontansprache. Das Sprachverstehen besteht dabei aus zwei Bereichen, die in das Design integriert werden: die Darstellung von Domänenwissen und der Mechanismus, der dieses Wissen nutzt, um die Äußerungen zu analysieren. Die Äußerungen müssen interpretiert werden können, auch wenn ihre Struktur keinem grammatikalisch korrekten Satz entspricht. Die Analyse findet daher nicht wie sonst üblich anhand syntaktischer Merkmale statt, sondern ist allein vom semantischen Inhalt gesteuert. Für eine möglichst freie Kommunikation zwischen Mensch und Roboter und den besonderen Anforderungen der situierten Spontansprache ist ein möglichst umfangreiches Wissen über die möglichen Dialoginhalte notwendig, das ebenfalls durch einen geeigneten Repräsentationsformalismus im System abgebildet werden muss. Die Bereitstellung des relevanten Wissens stellt daher ebenfalls einen zentralen Bereich des Sprachverstehens dar.

Das in dieser Arbeit beschriebene Konzept des Sprachverstehens wurde unter Berücksichtigung der besonderen Anforderungen der Mensch-Roboter-Kommunikation entwickelt. Sowohl die Wissensrepräsentation für das Verstehen der Äußerungen, die situierten semantischen Konzepte (SSUs), als auch der Verarbeitungsmechanismus, der auf dem Konzept der SSUs aufbaut, unterstützen die besondere Art der Dialoge.

Die SSUs stellen einen Repräsentationsformalismus für gesprochene Sprache dar. Sie bilden das semantische Wissen des Korpus für die Dialoge zwischen Mensch und Roboter ab. Ebenso stellen sie hierarchische Informationen zur Verfügung, die sowohl für den Verarbeitungsprozess als auch für die Bereitstellung der Dialogakte genutzt werden können. Zusätzlich kennzeichnen sie unterschiedlich wichtige Informationen und zusätzliche Informationen. Dieses Wissen wird genutzt, um die semantische Kohärenz und somit indirekt die Güte der Spracherkennungsleistung zu bewerten und um auf fehlende Informationen (ggf. aus der Szene) aufmerksam zu machen.

Das Konzept der SSUs ist speziell für die Integration zwischen Sprache und Szene konzipiert. Beispielsweise beinhalten die SSUs, die Objektinformationen bereitstellen, die Informationen, die zum Erkennen der Objekte relevant sind. Zusätzlich wurde eine SSU konzipiert, die auf mögliche Gesten hinweist. Der Repräsentationsformalismus unterstützt demnach das Konzept der Multimodalität des Robotersystems. Es gibt darüberhinaus SSUs, die generell für eine möglichst freie Interaktion mit dem Roboter zur Verfügung stehen und verschiedenste Äußerungsinhalte abbilden, wie z. B. Aufgaben, Fragen an den Roboter, Beschreibungen von Objekten oder Sachverhalten. Zusätzlich geben sie auch Hilfestellung, wie der Interaktionspartner kommuniziert und wie der Roboter darauf reagieren und sich ggf. an den Kommunikationsstil anpassen kann (z. B. durch die Darstellung von Höflichkeitsformeln). Einige SSUs wurden für spezielle Kontexte erstellt, die die gesamte Interaktionsfähigkeit des Robotersystems erhöhen, wie z. B. besondere SSUs zur Markierung von Informationen, die direkt die Themendetektion betreffen. Insgesamt umfasst der Sprachwortschatz 1400 Wörter (in Vollform), die auf etwa 150 SSUs abgebildet werden.

Der Verarbeitungsmechanismus verwendet neben dem Lexikon die SSUs, um eine semantische Analyse der Äußerungen zu erstellen. So wie auch die SSUs konzipiert sind, situierte Spontansprache möglichst gut abzubilden, so ist auch der Verarbeitungsmechanismus darauf ausgerichtet, diese besondere Form der Sprache zu verarbeiten und die Kommunikation zwischen Mensch und Roboter zu stützen. Da in diesem Kontext Äußerungen selten grammatikalischen Regeln folgen und daher nur schwerlich durch diese beschrieben werden können, werden die semantischen Beziehungen zwischen den Wörtern einer Äußerung genutzt, um eine semantische Interpretation zu gewinnen, die anschließend an den Dialogmanager weitergereicht wird, der die Interaktion zwischen dem Benutzer und dem Robotersystem steuert. Mit Hilfe einer Bewertungsfunktion wird das beste Ergebnis oder die besten Ergebnisse herausgefiltert, letzteres kommt bei der Verwendung von Homonymen in seltenen Fällen vor. Zusätzlich kann die Güte der semantischen Interpretation Aufschluss über die Erkennungsleistung der Spracherkennung geben. Die Bewertung gibt indirekt Hinweise, ob die Äußerung semantisch einen Sinn ergibt und sie somit auch aller Wahrscheinlichkeit nach so geäußert wurde oder ob möglicherweise Teile falsch erkannt wurden und somit der Sinn fehlt. Gleichzeitig wird für die Interpretation der Äußerung ein heuristischer Ansatz gewählt, um die Komplexität der möglichen Ableitungsbäume zu reduzieren. Der Verarbeitungsmechanismus kann somit in Echtzeit-Anwendungen wie mobilen Robotersystemen einsetzt werden. Das Problem der vollständigen Suche aller Ableitungen ist exponentiell und kann durch den Einsatz einer geeigneten Heuristik auf eine quadratische Funktion eingeschränkt werden.

Die Trennung der Verarbeitung vom Domänenwissen unterstützt ein offenes Architekturkonzept. So kann die Wissensdatenbank ausgetauscht, erweitert oder geändert werden, ohne dass das System selbst verändert werden muss. Der Austausch von verschiedenen Kontexten und auch Sprachen, wie bereits bei Deutsch und Englisch vollzogen, ist somit problemlos durchführbar. Der Datenaustausch in XML unterstützt die Offenheit der Architektur ebenfalls, da sie einen Datenaustausch unabhängig von der Programmiersprache ermöglicht.

Mit der Evaluierung des Mechanismus wurden zwei Bereiche überprüft. Zum einen wurden die Fähigkeiten und die Grenzen des Mechanismus dargestellt und die gute Eignung für die Verarbeitung situierter Spontansprache aufgezeigt. Zum anderen wurde der im Robotersystem BIRON integrierte Ansatz auf die Funktionalität im laufenden Betrieb überprüft. Hier zeigt sich ebenfalls, dass das Verfahren ein geeignetes Mittel darstellt, den Dialog mit Äußerungs-Interpretationen zu unterstützen: durch Informationen über den semantischen Inhalt, durch Angabe der Dialogakte, durch Hinweise auf Szeneinformationen und Informationen über die Güte der Erkennungsleistung (Hinweise auf Rückfragemöglichkeiten).

Zusammenfassend läßt sich sagen, dass der in dieser Arbeit vorgestellte Ansatz besonders geeignet für den Einsatz in Dialogsystemen mobiler Roboter ist. Das Verfahren kann in Echtzeit situierte Spontansprache verarbeiten und daraus die relevanten semantischen und soweit möglich auch pragmatischen Informationen bereitstellen. Ebenso kann es mit fehlerhaften Eingaben aus der Spracherkennung umgehen und sie verarbeiten. Es unterstützt den Dialogprozess durch zusätzliche Informationen über den Dialogakt und gibt Hilfestellung für die Verbindung von Sprache mit den visuellen Informationen aus der Szene. Darüber hinaus bietet das Konzept Hinweise auf die Güte der Spracherkennung. Insgesamt stellt es Wissen für die mögliche und sinnvolle Reaktion des gesamten Robotersystems bereit. Das Konzept unterstützt die Idee des offenen Architekturkonzeptes, indem die Wissensdatenbanken ebenfalls flexibel erweitert oder ausgetauscht werden können, ohne den Verarbeitungsmechanismus in irgendeiner Weise ändern zu müssen.

Das im Rahmen dieser Arbeit entwickelte Konzept zum Sprachverstehen verleiht einem mobilen Robotersystem die Fähigkeit, spontansprachliche und situierte Äußerungen zu verstehen und mit anderen Modalitäten in Beziehung zu setzen. Das Robotersystem erlangt mit Einsatz dieses Mechanismus tiefgehende kommunikative Fähigkeiten, die den Grundstock zur Verständigung in freien Dialogsituationen bilden. Dabei geht das Sprachverstehen in seiner Vollständigkeit weit über bereits existierende Sprachverstehenssysteme im Kontext mobiler Roboter hinaus.

Literaturverzeichnis

[All96] J. F. Allen, B. W. Miller, E. K. Ringger, T. Sikorski: *A Robust System for Natural Spoken Dialogue*, in *Proc. of the Annual Meeting of the Association for Computational Linguistucs (ACL'96)*, Morgan Kaufmann Publishers, 1996, S. 62–70.

[And99] M. Andersson, A. Orebäck, M. Lindstrom, H. I. Christensen: *ISR: An Intelligent Service Robot*, in H. I. Christensen, H. Bunke, H. Noltmeier (Hrsg.): *Sensor Based Intelligent Robots; International Workshop Dagstuhl Castle, Germany, September 28 – October 2, 1998. Selected Papers*, Bd. 1724 von *Lecture Notes in Computer Science*, Springer-Verlag, New York, NY, 1999, S. 287–310.

[Asf00] T. Asfour, K. Berns, R. Dillmann: *The Humanoid Robot ARMAR: Design and Control*, in *Proc. IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, MIT Press, Boston, MA, Sep. 2000.

[Aso01] H. Asoh, Y. Motomura, F. Asano, I. Hara, S. Hayamizu, K. Itou, T. Kurita, T. Matsui, N. Vlassis, R. Bunschoten, B. Kröse: *Jijo-2: An Office Robot that Communicates and Learns*, *IEEE Intelligent Systems*, Bd. 16, Nr. 5, 2001, S. 46–55.

[Aus62] J. L. Austin: *How to do things with words*, Oxford: Oxford University Press, 1962.

[Aus85] J. L. Austin: *Zur Theorie der Sprechakte (How to do things with words)*, Ditzingen: Reclam, 1985.

[Bak98] C. F. Baker, C. J. Fillmore, J. B. Lowe: *The Berkeley FrameNet project*, in *Proc. of the Int. Conf. on Computational Linguistics (COLING/ACL)*, Montreal, Canada, 1998.

[Bar68] R. G. Barker: *Ecological psychology*, Stanford University Press, Stanford, USA, 1968.

[Bar87] G. E. Barton, R. C. Berwick, E. S. Ristad: *Computational complexity and Natural Language*, MIT Press, Cambridge, MA, USA, 1987.

[Bar01] C. Bartneck, M. Okada: *Robotic User Interfaces*, in *Proc. Int. Conf. on Human and Computer*, Aizu, Japan, Sep. 2001, S. 130–140.

[Bau01] C. Bauckhage, G. A. Fink, J. Fritsch, F. Kummert, F. Lömker, G. Sagerer, S. Wachsmuth: *An Integrated System for Cooperative Man-Machine Interaction*, in *IEEE*

International Symposium on Computational Intelligence in Robotics and Automation, Banff, Canada, 2001, S. 328–333.

[Bis99] R. Bischoff, T. Jain: *Natural Communication and Interaction with Humanoid Robots*, in *Proc. Int. Symp. on Humanoid Robots (HURO)*, Tokyo, Japan, Okt. 1999, S. 121–128.

[Bis02a] R. Bischoff, V. Graefe: *Demonstrating the Humanoid Robot HERMES at an Exhibition: A Long-Term Dependability Test*, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems; Workshop on Robots at Exhibitions*, Lausanne, Switzerland, 2002.

[Bis02b] R. Bischoff, V. Graefe: *Dependable Multimodal Communication and Interaction with Robotic Assistants*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, Berlin, Germany, Sep. 2002, S. 300–305.

[Blo95] H. U. Block, S. Schachtl: *What a grammar of spoken dialogues has to deal with*, in G. Heyer, H. Haugeneder (Hrsg.): *Language Engineering, Advanced Studies in Computer Science*, Vieweg, Braunschweig, 1995, S. 101–126.

[Blo00] H. U. Block, T. Ruland: *Integrated Shallow Linguistic Processing*, in W. Wahlster (Hrsg.): *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag, Berlin, Heidelberg, 2000, S. 143–162.

[Böh98] H.-J. Böhme, U.-D. Braumann, A. Brakensiek, A. Corradini, M. Krabbes, H.-M. Gross: *User Localisation for Visually-based Human-Machine-Interaction*, in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, IEEE Press, Nara, Japan, April 1998, S. 486–491.

[Böh03] H.-J. Böhme, T. Wilhelm, J. Key, C. Schauer, C. Schröter, H.-M. Groß, T. Hempel: *An Approach to Multi-modal Human-Machine Interaction for Intelligent Service Robots*, *Robotics and Autonomous Systems*, Bd. 44, Nr. 1, Juli 2003, S. 83–96.

[Bor91] R. D. Borsley: *Syntactic Theory – A Unified Approach*, Edward Arnold, London, New York, Sydney, Auckland, 1991.

[Bos02] J. Bos: *Compilation of Unification Grammars with Compositional Semantics to Speech Recognition Packages*, in *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, 2002, S. 106–112.

[BP99a] H. Brand-Pook: *Eine Sprachverstehenskomponente in einem Konstruktionsszenario*, Dissertation, Universität Bielefeld, Technische Fakultät, Angewandte Informatik, Bielefeld, Deutschland, Mai 1999.

[BP99b] H. Brand-Pook, G. A. Fink, S. Wachsmuth, G. Sagerer: *Integrated Recognition and Interpretation of Speech for a Construction Task Domain*, in *Proc. 8th Int. Conf. on Human-Computer Interaction (CHI'99)*, Bd. 1, München, 1999, S. 550–554.

[Bra99] P. Brandt, D. Dettmer, G. S. R.-A-Dietrich: *Sprachwissenschaft: ein roter Faden für das Studium*, Böhlau Verlag, Köln, Weimar, Wien, Böhlau, 1999.

[Bre99] C. Breazeal, B. Scassellati: *A Context-Dependent Attention System for a Social Robot*, in T. Dean (Hrsg.): *Proc. Int. Joint Conf. on Artificial Intelligence*, Bd. 2, Morgan Kaufmann Publishers Inc., Stockholm, Sweden, Juli/Aug. 1999, S. 1146—1151.

[Bre00] C. Breazeal, A. Edsinger, P. Fitzpatrick, B. Scassellati, P. Varchavskaia: *Social Constraints on Animate Vision*, *IEEE Intelligent Systems*, Bd. 15, Nr. 4, Juli/Aug. 2000, S. 32–37.

[Bre01] J. Bresnan: *Lexcial-Functional Syntax*, Blackwell Publishers, Oxford, 2001.

[Bre04] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, D. Mulanda: *Humanoid Robots as Cooperative Partners for People*, *Int. Journal of Humanoid Robots*, 2004.

[Bro83] D. K. Brotz: *Message System Mores: Etiquette in Laurel*, *ACM Transaction on Office Information Systems*, Bd. 1, 1983, S. 179–192.

[Bro90] R. A. Brooks: *Elephants Don't Play Chess*, *Robotics and Autonomous Systems*, Bd. 6, Nr. 1–2, Juni 1990, S. 3–15.

[Bro99] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, M. M. Williamson: *The Cog Project: Building a Humanoid Robot*, in C. L. Nehaniv (Hrsg.): *Computation for Metaphors, Analogy, and Agents*, Bd. 1562 von *Lecture Notes in Computer Science*, Springer-Verlag, New York, NY, 1999, S. 52–87.

[Bru66] J. S. Brunder: *Towards a Theory of Instruction*, Norton, New York, 1966.

[Bru75] B. Bruce: *Case Systems for Natural Language*, *Artificial Intelligence*, Bd. 6, 1975, S. 327–360.

[Bur98] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun: *The Interactive Museum Tour-Guide Robot*, in *Proc. Nat. Conf. on Artificial Intelligence (AAAI)*, AAAI/MIT Press, Madison, WI, Juli 1998, S. 11–18.

[Bur99] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun: *Experiences with an Interactive Museum Tour-Guide Robot*, *Artificial Intelligence*, Bd. 114, Nr. 1–2, Okt. 1999, S. 3–55.

[Cap92] B. Caprile, G. Lazzari, L. Stringa: *Autonomous navigation and speech in mobile robot of MAIA*, in *OE/Technology '92*, The International Society for Optical Engineering, Boston, MA, USA, 1992, S. 15–21.

[Car70] J. R. Carbonell: *AI in CAI: An artificial intelligent approach to computer-assisted instruction*, *IEEE transaction on Man Machine System*, Bd. 11, Nr. 4, 1970, S. 190–202.

[Cas00] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, H. Yan: *Human Conversation as a System Framework: Designing Embodied Conversational Agents*, in J. Cassell, J. Sullivan, S. Prevost, E. Churchill (Hrsg.): *Embodied Conversational Agents*, Kap. 2, MIT Press, Cambridge, MA, 2000, S. 29–63.

[Cho72] N. Chomsky (Hrsg.): *Studies on Semantics in Generative Grammar*, The Hague: Mouton, Berlin and New York, 1972.

[Cho81] N. Chomsky: *Lectures on government and binding*, Foris Publications, reprint., 7th edition. berlin and new york: mouton de gruyter, 1993. Ausg., 1981.

[Cla97] W. J. Clancey: *Situated Cognition. On Human Knowledge and Computer Representations*, New York: Cambridge University Press, 1997.

[Col69] L. S. Coles: *Talking with a Robot in English*, in *Proc. Int. Joint Conf. on Artificial Intelligence*, Washington D.C., 1969, S. 587–596.

[Con92] J. Connell: *SSS: A hybrid architecture applied to robot navigation*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, Bd. 3, Nice, France, 1992, S. 2719–2724.

[Dau04] K. Dautenhahn: *Robots We Like to Live With?! – A Developmental Perspective on a Personalized, Life-Long Robot Companion*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, Kurashiki, Okayama Japan, Sep. 2004, S. 17–22.

[Dik91] S. C. Dik: *Functional Grammar*, in F. G. Droste, J. E. Joseph (Hrsg.): *Linguistic theory and grammatical Description*, John Benjamins, Amsterdam, Philadelphia, 1991, S. 247–274.

[Doh94] W. Dohmen: *Kooperative Systeme - Techniken und Chancen*, München Wien: Carl Hanser Verlag, 1994.

[Doi02] M. Doi, K. Suzuki, S. Hashimoto: *Integrated Communicative Robot "BUGNOID"*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, Berlin, Germany, Sep. 2002, S. 259–265.

[Dor87] F. Dorsch, H. Häcker, K.-H. Stapf: *Dorsch Psychologisches Wörterbuch*, Bern: Hans Huber, 1987.

[Dow93] J. Dowding, J. M. Gawron, D. E. Appelt, J. Bear, L. Cherny, R. Moore, D. B. Moran: *GEMINI: A Natural Language System for Spoken-Language Understanding*, in *Proc. of the ACL*, 1993, S. 54–61.

[Ehr02] M. Ehrenmann, R. Becher, B. Giesler, R. Zöllner, O. Rogalla, R. Dillmann: *Interaction with Robot Assistants: Commanding ALBERT*, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Lausanne, Schweiz, Sep. 2002.

[Eli02] P. Elinas, J. Hoey, D. Lahey, J. Montgomery, D. Murray, S. Se, J. J. Little: *Waiting with José, a vision-based mobile robot*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, Bd. 4, IEEE Press, Washington DC, USA, Mai 2002, S. 3698–3705.

[Erk03] K. Erk, A. Kowalski, S. Pado, M. Pinkal: *Building a Resource for Lexical Semantics*, in *Proc. of the 17th International Conference of Linguists (CIL), Prague.*, 2003.

[Fel99] C. Fellbaum (Hrsg.): *WordNet : an electronic lexical database*, MIT Press, Cambridge, Mass, 1999.

[Fil68] C. J. Fillmore: *The Case for Case*, in E. Bach, R. Harms (Hrsg.): *Universals in Linguistic Theory*, Holt, Rinehart, and Winston, New York, 1968, S. 1–90.

[Fil76] C. J. Fillmore: *Frame semantics and the nature of language*, in *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Bd. 280, 1976, S. 20–32.

[Fil01] C. J. Fillmore, C. F. Baker: *Frame Semantics for Text Understanding*, in *Proc. of WordNet and Other Lexical Recources Workshop*, Juni 2001.

[Fin95] G. A. Fink, N. Jungclaus, H. Ritter, G. Sagerer: *A Communication Framework for Heterogeneous Distributed Pattern Analysis*, in V. L. Narasimhan (Hrsg.): *Proc. Int. Conf. on Algorithms and Architectures for Parallel Processing*, IEEE Press, Brisbane, Australia, April 1995, S. 881–890.

[Fis95] K. Fischer, M. Johanntokrax: *Ein linguistisches Merkmalsmodell für die Lexikalisierung von diskurssteuernden Partikeln*, 18/95, Situierte Künstliche Kommunikatoren, SFB 360, Universität Bielefeld, 1995.

[Fis96] K. Fischer, B. Wrede, C. Brindöpke, M. Johanntokrax: *Quantitative und funktionale Analysen von Diskurspartikeln im Computer Talk*, *Sprache und Datenverarbeitung. International Journal of Language Data Processing*, 1996, S. 85–100.

[Flo97] C. Floyd, A. Krabbel, S. Ratusky, I. Wetzel: *Zur Evolution der evolutionären Systementwicklung: Erfahrungen aus einem Krankenhausprojekt*, *Informatik Spektrum*, Bd. 20, 1997, S. 13–20.

[Fon01] T. W. Fong, C. Thorpe, C. Baur: *Collaboration, Dialogue, and Human-Robot Interaction*, in *Proc. Int. Symp. on Robotics Research*, Springer-Verlag, Lorne, Victoria, Australia, Nov. 2001.

[Fon03] T. W. Fong, C. Thorpe, C. Baur: *Robot, Asker of Questions*, *Robotics and Autonomous Systems*, Bd. 42, Nr. 3–4, März 2003, S. 235–243.

[Fri05] J. Fritsch, M. Kleinehagenbrock, A. Haasch, S. Wrede, G. Sagerer: *A Flexible Infrastructure for the Development of a Robot Companion with Extensible HRI-Capabilities*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, IEEE Press, Barcelona, Spain, April 2005, S. 3419–3425.

[Fry98] J. Fry, H. Asoh, T. Matsui: *Natural Dialogue with the Jijo-2 Office Robot*, in *Proc. of the 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Bd. 4, 1998, S. 1278–1283.

[Gaz85] G. Gazdar, E. Klein, I. Sag: *Generalized Phrase Structure Grammar*, Harvard University Press, Cambridge, MA, 1985.

[Gil00] D. Gildea, D. Jurafsky: *Automatic Labeling of Semantic Roles*, in *Proc. of the ACL*, Hong Kong, Okt. 2000, S. 512–520.

[Goe96] K. U. Goecke, J.-T. Milde, H. Lobin: *Aufgabenorientierte Verarbeitung von Interventionen und Instruktionen*, Report 96/7, Situierte Künstliche Kommunikatoren, SFB 360, Universität Bielefeld, 1996.

[Gra04] B. Graf, M. Hans, R. D. Schraft: *Care-O-bot II—Development of a Next Generation Robotic Home Assistant*, *Autonomous Robots*, Bd. 16, Nr. 2, März 2004, S. 193–205.

[Gro86] B. J. Grosz, C. L. Sidner: *Attention, Intention, and the Structure of Discourse*, *Computational Linguistics*, Bd. 12, Nr. 3, 1986, S. 175–204.

[Gru67] J. S. Gruber: *Studies und lexical relations*, Bloomington, 1967.

[Gug99] E. Guglielmelli, C. Laschi, P. Dario: *Robots for Personal Use: Humanoids vs. Distributed Systems*, in *Proc. Int. Symp. on Humanoid Robots (HURO)*, Tokyo, Japan, Okt. 1999.

[Haa04] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, G. Sagerer: *BIRON – The Bielefeld Robot Companion*, in E. Prassler, G. Lawitzky, P. Fiorini, M. Hägele (Hrsg.): *Proc. Int. Workshop on Advances in Service Robotics*, Fraunhofer IRB Verlag, Stuttgart, Germany, Mai 2004, S. 27–32.

[Haa05] A. Haasch, N. Hofemann, J. Fritsch, G. Sagerer: *A Multi-Modal Object Attention System for a Mobile Robot*, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, IEEE Press, Edmonton, Alberta, Canada, Aug. 2005, S. 1499–1504.

[Har01] I. Haritaoglu, A. Cozzi, D. Koons, M. Flickner, D. Zotkin, Y. Yacoob: *Attentive Toys*, in *Proc. IEEE Int. Conf. on Multimedia and Expo*, IEEE Press, Tokyo, Japan, Aug. 2001, S. 1124–1127.

[Has02] S. Hashimoto, S. Narita, H. Kasahara, K. Shirai, T. Kobayashi, A. Takanishi, S. Sugano, J. Yamaguchi, H. Sawada, H. Takanobu, K. Shibuya, T. Morita, T. Kurata, N. Onoe, K. Ouchi, T. Noguchi, Y. Niwa, S. Nagayama, H. Tabayashi, I. Matsui, M. Obata, H. Matsuzaki, A. Murasugi, T. Kobayashi, S. Haruyama, T. Okada, Y. Hidaki, Y. Taguchi, K. Hoashi, E. Morikawa, Y. Iwano, D. Araki, J. Suzuki, M. Yokoyama, I. Dawa, D. Nishino, S. Inoue, T. Hirano, E. Soga, S. Gen, T. Yanada, K. Kato, S. Sakamoto, Y. Ishii, S. Matsuo, Y. Yamamoto, K. Sato, T. Hagiwara, T. Ueda,

N. Honda, K. Hashimoto, T. Hanamoto, S. Kayaba, T. Kojima, H. Iwata, H. Kubodera, R. Matsuki, T. Nakajima, K. Nitto, D. Yamamoto, Y. Kamizaki, S. Nagaike, Y. Kunitake, S. Morita: *Humanoid Robots in Waseda University—Hadaly-2 and WABIAN*, *Autonomous Robots*, Bd. 12, Nr. 1, Jan. 2002, S. 25–38.

[Hir94] J. Hirschberg, C. H. Nakatani: *A Corpus-based study of repair cues in spontaneous speech*, *Acoustical Society of America*, Bd. 95, Nr. 3, März 1994, S. 1603–1616.

[Hir98] K. Hirai, M. Hirose, Y. Haikawa, T. Takenaka: *The Development of Honda Humanoid Robot*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, Bd. 2, IEEE Press, Leuven, Belgium, Mai 1998, S. 1321–1326.

[Hof04] N. Hofemann, J. Fritsch, G. Sagerer: *Recognition of Deictic Gestures with Context*, in C. E. Rasmussen, H. H. Bülthoff, M. A. Giese, B. Schölkopf (Hrsg.): *Pattern Recognition; 26th DAGM Symposium, Tübingen, Germany, August/September 2004. Proceedings*, Bd. 3175 von *Lecture Notes in Computer Science*, Springer-Verlag, Heidelberg, Germany, 2004, S. 334–341.

[Hüt03] H. Hüttenrauch, A. Green, K. Severinson-Eklundh, L. Oestreicher, M. Norman: *Involving Users in the Design of a Mobile Office Robot*, *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 2003.

[Hüw04] S. Hüwel, F. Kummert: *Interpretation of Situated Human-Robot Dialogues*, in *Proc. of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, Birmingham, UK, Jan. 2004, S. 120–125.

[Hüw05] S. Hüwel, B. Wrede, G. Sagerer: *Semantisches Parsing mit Frames für robuste multimodale Mensch-Maschine Kommunikation*, in *GLDV-conference 2005*, Bonn, Germany, März/April 2005.

[Hüw06a] S. Hüwel, B. Wrede: *Situated Speech Understanding for Robust Multi-Modal Human-Robot Communication*, in *Proc. of the International Conference on Computational Linguistics (COLING/ACL)*, ACL Press, 2006.

[Hüw06b] S. Hüwel, B. Wrede, G. Sagerer: *Robust Speech Understanding for Multi-Modal Human-Robot Communication*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, 2006.

[Ios02] I. Iossifidis, C. Bruckhoff, C. Theis, C. Grote, C. Faubel, G. Schoener: *Cora: An Anthromorphic Robot assistant for Human Environment*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, Berlin, Germany, Sep. 2002, S. 392–398.

[Iss93] S. Issar, W. Ward: *CMU's Robust Spoken Language Understanding System*, in *Proc. Europ. Conf. on Speech Communication and Technology (Eurospeech)*, 1993, S. 2147–2150.

[Jac72] R. S. Jackendoff: *Semantic Interpretation in Generative Grammar*, Cambridge Mass, Cambridge, Ma, 1972.

[Jac83] R. S. Jackendoff: *Semantics and Cognition*, MIT Press, Cambridge, MA, 1983.

[Jac90] R. S. Jackendoff: *Semantic Structures*, MIT Press, 1990.

[Joh04] M. Johnston, S. Bangalore: *Robust Multimodal Understanding*, in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, 2004.

[Kam97] C. A. Kamm, M. A. Walker: *Design and Evaluation of Spoken Dialog Systems*, in *Proc. of the ASRU Workshop*, IEEE, Santa Barbara, CA, USA, 1997, S. 11–18.

[Kan04] K. Kaneko, F. Kanehiro, S. Kajita, H. Hirukawa, T. Kawasaki, M. Hirata, K. Akachi, T. Isozumi: *Humanoid Robot HRP-2*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, Bd. 2, IEEE Press, New Orleans, LA, April/Mai 2004, S. 1083–1090.

[Kap04] F. Kaplan, V. V. Hafner: *The Challenge of Joint Attention*, in *Proc. of the Fourth International Workshop on Epigenetic Robotics*, 2004, S. 67–74.

[Kat63] J. J. Katz, J. A. Fodor: *The structure of a Semantic Theory*, *Language*, Bd. 39, 1963, S. 170–210.

[Kim04] G. Kim, W. Chung, S. Han, K.-R. Kim, M. Kim, R. H. Shinn: *The Autonomous Tour-Guide Robot Jinny*, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Bd. 3, Sendai, Japan, Sep./Okt. 2004, S. 3450–3455.

[Kle04] M. Kleinehagenbrock: *Interaktive Verhaltenssteuerung für Robot Companions*, Dissertation, Universität Bielefeld, Technische Fakultät, Angewandte Informatik, Bielefeld, Deutschland, Dez. 2004.

[Koi00] H. Koike, Y. Sato, Y. Kobayashi, H. Tobita, M. Kobayashi: *Interactive Textbook and Interactive Venn Diagramm: Natural and Intuitive Interfaces on Augmented Desk System*, in *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, 2000, S. 121–128.

[Kom95] R. Kompe, W. Eckert, A. Kiessling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini: *Towards domain-independent understanding of spontaneous speech*, in *Proc. of the European Congress on Intelligent Techniques and Soft Computing*, Bd. 3, Aachen, Germany, Aug. 1995, S. 2315–3319.

[Kop03] S. Kopp, B. Jung, N. Leßmann, I. Wachsmuth: *Max – A Multimodal Assistant in Virtual Reality Construction*, *KI - Künstliche Intelligenz, Special issue on Embodied Conversational Agents*, Bd. 17, Nr. 4, 2003, S. 11–17.

[Kop05] S. Kopp, L. Gesellensetter, N. Krämer, I. Wachsmuth: *A Conversational Agent as Museum Guide – design and evaluation of a Real-World Application*, in Panayiotopoulos, others (Hrsg.): *Intelligent Virtual Agents*, LNAI 3661, Springer, Berlin, 2005, S. 329–343.

[Kra92] J. Krause: *Natürlichsprachliche Mensch-Computer-Interaktion als technisierte Kommunikation: Die computer talk-Hypothese*, in J. Krause, L. Hitzenberger (Hrsg.): *Computer Talk*, Olms, Hildesheim, Germany, 1992, S. 1–29.

[Kri01] S. Kristensen, S. Horstmann, J. Klandt, F. Lohnert, A. Stopp: *Human-Friendly Interaction for Learning and Cooperation*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, Bd. 3, IEEE Press, Seoul, Korea, Mai 2001, S. 2590–2595.

[Kro00] S. Kronenberg: *Cooperation in Human-Computer Communication*, Dissertation, Universität Bielefeld, Technische Fakultät, Angewandte Informatik, Bielefeld, Deutschland, 2000.

[Krö03] B. J. A. Kröse, J. M. Porta, A. J. N. van Breemen, K. Crucq, M. Nuttin, E. Demeester: *Lino, the User-Interface Robot*, in E. H. L. Aarts, R. Collier, E. van Loenen, B. E. R. de Ruyter (Hrsg.): *Ambient Intelligence; First European Symposium, EUSAI 2003, Veldhoven, The Netherlands, November 3–4, 2003. Proceedings*, Bd. 2875 von *Lecture Notes in Computer Science*, Springer-Verlag, Heidelberg, Germany, 2003, S. 264–274.

[Lae95] T. Laengle, T. C. Lueth, E. Stopp, G. Herzog, G. Kamstrup: *KANTRA - A Natural Language Interface for Intelligent Robots*, 114, DFKI, Saarbrücken, Deutschland, März 1995.

[Lan03] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, G. Sagerer: *Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot*, in *Proc. Int. Conf. on Multimodal Interfaces*, ACM, Vancouver, Canada, Nov. 2003, S. 28–35.

[Lau01] S. Lauria, G. Bugmann, T. Kyriacou, J. Bos, E. Klein: *Personal Robot Training via Natural Language Instructions*, *IEEE Intelligent Systems, 16*, Bd. 16, Nr. 3, Sep./Okt. 2001, S. 38–45.

[Lau02] S. Lauria, T. Kyriacou, G. Bugmann, J. Bos, E. Klein: *Converting Natural Language Route Instructions into Robot-Executable Procedures*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, Berlin, Germany, 2002, S. 223–228.

[Lav96] A. Lavie: *GLR*: A Robust Grammar-Focused Parser for Spontaneously Spoken Language*, Dissertation, Carnegie Mellon University, Pittsburgh, PA, Mai 1996.

[Lav97] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, P. Zhan: *JanusIII: speech-to-speech translation in multiple languages*, in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, IEEE, Munich, Germany, 1997, S. 99–102.

[Lay01] K. Lay, E. Prassler, R. Dillmann, G. Grunwald, M. Hägele, G. Lawitzky, A. Stopp, W. von Seelen: *MORPHA: Communication and Interaction with Intelligent, Anthropomorphic Robot Assistants*, in *Proc. Int. Status Conf. of the German Lead Projects in Human-Computer-Interaction*, Saarbrücken, Germany, Okt. 2001, S. 67–77.

[Lem01a] O. Lemon, A. Bracy, A. Gruenstein, S. Peters: *A Multi-Modal Dialogue System for Human Robot Conversation*, in *Proc. North American Chapter of the ACL*, Pittsburgh, USA, Juni 2001.

[Lem01b] O. Lemon, A. Bracy, A. Gruenstein, S. Peters: *The WITAS Multi-Modal Dialogue System I*, in *Proc. Europ. Conf. on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, Sep. 2001, S. 1559–1562.

[Lev83] W. J. Levelt: *Monitoring and self-repairs in speech*, *Cognition*, Bd. 14, 1983, S. 41–104.

[Lew95] J. P. Lewis: *Fast Template matching*, in *Proc. Int. Conf. on Vision Interface*, Quebec, Canada, 1995, S. 120–123.

[Li04] S. Li, M. Kleinehagenbrock, J. Fritsch, B. Wrede, G. Sagerer: *"BIRON, let me show you something": Evaluating the Interaction with a Robot Companion*, in W. Thissen, P. Wieringa, M. Pantic, M. Ludema (Hrsg.): *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, Special Session on Human-Robot Interaction*, IEEE, The Hague, The Netherlands, Okt. 2004, S. 2827–2834.

[Li06a] S. Li, B. Wrede, G. Sagerer: *A computational model of multi-modal grounding*, in *Proc. ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006*, ACL Press, 2006.

[Li06b] S. Li, B. Wrede, G. Sagerer: *A dialog system for comparative user studies on robot verbal behavior*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, 2006.

[Lim00] H.-O. Lim, A. Takanishi: *Waseda Biped Humanoid Robots Realizing Human-like Motion*, in *Proc. Int. Workshop on Advanced Motion Control*, IEEE Press, Nagoya, Japan, März/April 2000, S. 525–530.

[Lob93a] H. Lobin: *Koordinationssyntax als prozedurales Phänomen*, Narr, Tübingen, 1993.

[Lob93b] H. Lobin: *Situiertheit*, *KI*, Bd. 1/93, Nr. 61, 1993.

[Lob98] H. Lobin: *Handlungsanweisungen: sprachliche Spezifikation teilautonomer Aktivität*, Wiesbaden: Deutscher Universitäts Verlag, 1998.

[Lop03a] L. S. Lopes, A. Teixeira, M.Rodrigues, D. C.Teixeira, L. Feirreira, P. Soares, J.Giro, N. Snica: *Towards a Personal Robot with Language Interface*, in *Proc. Europ. Conf. on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, 2003, S. 1559–1562.

[Lop03b] L. S. Lopes, A. Teixeira, M. Rodrigues, D. Gomes, J. Girão, C. Teixeira, N. Sénica, L. Ferreira, P. Soares: *A Robot with Natural Interaction Capabilities*, in *Proc. IEEE Int'l Conf. on Emerging Technologies and Factory Automation*, Bd. 1, 2003, S. 605–612.

[Lop05] L. S. Lopes, A. Teixeira, M. Quindere, M. Rodrigues: *From Robust Spoken Language Understanding to Knowledge Aquisition and Management*, in *Proc. Europ. Conf. on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal, 2005.

[Lue94] T. C. Lueth, T. Laengle, G. Herzog, E. Stopp, U. Rembold: *KANTRA - Human-Machine Interaction for Intelligent Robots using Natural Language*, in *IEEE International Workshop on Robot and Human Communication*, Bd. 4, 1994, S. 106–110.

[Maa06] J. F. Maas, T. Spexard, J. Fritsch, B. Wrede, G. Sagerer: *BIRON, what's the topic? - A Multi-Modal Topic Tracker for improved Human-Robot Interaction*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE, September 2006.

[Mae94] P. Maes: *Agents that reduce work and information overload*, *Communications of the ACM*, Bd. 7, Nr. 34, 1994, S. 31–40.

[Mas94] M. Mast, F. Kummert, U. Ehrlich, G. A. Fink, T. Kuhn, H. Niemann, G. Sagerer: *A Speech Understanding and Dialog System with a Homogeneous Linguistic Knowledge Base*, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Bd. 16, Nr. 2, 1994, S. 179–194.

[Mat99] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, N. Otsu: *Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services*, in *Proc. Nat. Conf. on Artificial Intelligence (AAAI)*, AAAI/MIT Press, Orlando, FL, Juli 1999, S. 621–627.

[Mat01] Y. Matsusaka, S. Fujie, T. Kobayashi: *Modelling of Conversational Strategy for the Robot Participating in the Group Conversation*, in P. Dalsgaard, B. Lindberg, H. Benner, Z. Tan (Hrsg.): *Proc. Europ. Conf. on Speech Communication and Technology (Eurospeech)*, Bd. 3, Aalborg, Denmark, Sep. 2001, S. 2173–2176.

[McK98] D. McKelvie: *The Syntax of Disfluency in Spontaneous Spoken Language*, Research Paper HCRC/RP-95, HCRC, University of Edinburgh, Mai 1998.

[Mei99] J. Meibauer: *Pragmatik: eine Einführung*, Stauffenburg Verlag, 1999.

[Mel88] I. A. Mel'cuk: *Dependency Syntax: Theory and Practice*, State University of New York Press, 1988.

[Men02] W. Menzel: *Parsing mit inkonsistenten Grammatiken*, *Kognitionswissenschaft*, Bd. 9, Nr. 4, 2002, S. 175–184.

[Mil97] J. T. Milde, K. Peters, S. Strippgen: *Situated Communication with Robots*, in *First Int. Workshop on Human-Computer-Conversation*, Juli 1997.

[Min81] M. Minsky: *A Framework for Representing Knowledge*, in J. Haugeland (Hrsg.): *Mind Design*, MIT Press, Cambridge, 1981, S. 95–128.

[Min96] W. Minker: *A Stochastic Case Frame Approach for Natural Language Understanding*, in *Proc. Int. Conf. on Spoken Language Processing*, 1996.

[Mon02] M. Montemerlo, J. Pineau, N. Roy, S. Thrun, V. Verma: *Experiences with a Mobile Robotic Guide for the Elderly*, in *Proc. Nat. Conf. on Artificial Intelligence (AAAI)*, AAAI/MIT Press, Edmonton, AB, Juli 2002, S. 587–592.

[Mül99] S. Müller: *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche.*, Niemeyer, 1999.

[Nag04] Y. Nagai: *Understanding the Development of Joint Attention from a Viewpoint of Cognitive Developmental Robotics*, Dissertation, Osaka University, 2004.

[Nak01] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, H. Kitano: *Real-Time Auditory and Visual Multiple-Object Tracking for Humanoids*, in B. Nebel (Hrsg.): *Proc. Int. Joint Conf. on Artificial Intelligence*, Bd. 2, Morgan Kaufmann Publishers Inc., Seattle, WA, Aug. 2001, S. 1425–1432.

[Nas00] C. Nass, Y. Moon: *Machines and Mindlessness: Social Responses to Computers*, *Journal of Social Issues*, Bd. 56, Nr. 1, 2000, S. 81–103.

[Nig95] L. Nigay, J. Coutaz: *A Generic Platform for Addressing the Multimodal Challenge*, in *CHI*, Denver, USA, 1995, S. 98–105.

[Nil84] N. J. Nilsson: *Shakey the Robot*, Technical Note 323, AI Center, SRI International, Menlo Park, CA, 1984.

[Nor94] D. A. Norman: *How might people interact with agents*, *Communications of the ACM*, Bd. 7, Nr. 37, 1994, S. 68–71.

[Nou99] I. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, A. Soto: *An Affective Mobile Educator with a Full-time Job*, *Artificial Intelligence*, Bd. 114, Nr. 1–2, Okt. 1999, S. 95–124.

[Nöt89] E. Nöth: *Prosodische Information in der automatischen Spracherkennung - Berechnung und Anwendung*, Dissertation, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 1989.

[Nöt99] E. Nöth, M. Boros, J. Haas, V. Warnke, F. Gallwitz: *A Hybrid Approach To Spoken Dialogue Understanding: Prosody, Statistics And Partial Parsing*, 1999.

[Oku01] H. G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, H. Kitano: *Human-Robot Interaction through Real-Time Auditory and Visual Multiple-Talker Tracking*, in *Proc. of the 2001 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2001.

[Par01] H. K. Park, H. S. Hong, H. J. Kwon, M. J. Chung: *A Nursing Robot System for the Elderly and the Disabled*, in *Proc. Int. Workshop on Human-friendly Welfare Robotic Systems*, Daejeon, Korea, Jan. 2001, S. 122–126.

[Pau92] D. B. Paul, J. M. Baker: *The Design for the Wall Street Journal-based CSR Corpus*, in *Speech and Natural Language Workshop*, Morgan Kaufmann, 1992.

[PB02] U. Pankoke-Babatz: *Designkonzept für Systeme zur computergestützten Zusammenarbeit unter Nutzung der Behavior-Setting-Theorie*, Dissertation, Universität Dortmund, Fachbereich Informatik, 2002.

[Per99] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh: *Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy*, in *Proc. IEEE Int. Symp. on Computational Intelligence in Robotics and Automation (CIRA)*, IEEE Press, Monterey, CA, Nov. 1999, S. 208–213.

[Per01] D. Perzanowski, A. C. Schulz, W. Adams, E. Marsh, M. Bugajaska: *Building a Multimodal Human-Robot Interface*, *IEEE Journal on Intelligent Systems*, Feb. 2001, S. 16–21.

[Pet99] K. Peters: *Natürlichsprachliche Kommunikation mit handelnden Systemen*, Logos Verlag, Berlin, Germany, 1999.

[Pol94] C. Pollard, I. Sag: *Head-driven Phrase Structure Grammar*, University of Chicago Press, Chicago, 1994.

[Pot99] A. Potamianos, H. Kwo, C. Lee, A. Pargellis, A. Saad, Q. Zhou: *Design Principles and Tools for Multimodal Dialog Systems*, in *Proc. of ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, Kloster Irsee, Germany, Juni 1999, S. 141–144.

[Pri89] W. Prinz: *Computer Based Group Communication – the AMIGO Activity Model*, in U. Pankoke-Babatz (Hrsg.): *Survey of group communication models and systems*, Chichester: Ellis Horwood, 1989, S. 127–180.

[Qui68] M. R. Quillian: *Semantic Memory*, in M. Minsky (Hrsg.): *Semantic Information Processing*, MIT Press, Cambridge Mass., 1968, S. 227–270.

[Rog02] O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, R. Dillmann: *Using Gesture and Speech Control for Command a Robot Assistant*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, IEEE Press, Berlin, Germany, Sep. 2002, S. 454–459.

[Ros98] C. P. Rosé, A. Lavie: *LCFlex: An Efficient Robust Left-Corner Parser*, University of Pittsburgh, 1998.

[Ros00] C. P. Rosé: *A framework for robust semantic interpretation*, 2000.

[Roy00a] D. Roy: *Integration of speech and vision using mutual information*, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2000, S. 2369–2372.

[Roy00b] N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Magaritis, M. Montemerlo, J. Pineau, J. Schulte, S. Thrun: *Towards Personal Service Robots for the Elderly*, in *Proc. Int. Workshop on Interactive Robotics and Entertainment*, Pittsburgh, PA, April/Mai 2000.

[Rus95] S. Russel, P. Norvig: *Artificial Intelligence, a Modern Approach*, Prentice Hall, 1995.

[Sae97] J. Saeed: *Semantics*, Blackwell Publishers, 1997.

[Sch90] R. C. Schank, D. B. Leake: *Creativity and learning in a case based explainer*, *Machine learning: Paradigms and methods*, 1990, S. 353–385.

[Sch95] B. Schmitz, J. J. Quantz: *Dialogue Acts in automatic Dialogue Processing*, in *Proc. Sixth Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, Leuven, 1995, S. 33–47.

[Sch97] M. Schwarz, J. Chur: *Semantik. Ein Arbeitsbuch*, Narr, Tübingen, 1997.

[Sch99] C. Schaeffer, T. May: *Care-O-bot^{TM}: A System for Assisting Elderly or Disabled Persons in Home Environments*, in C. Bühler, H. Knops (Hrsg.): *Assistive Technology on the Threshold of the New Millennium; AAATE 99, 5th European Conference for the Advancement of Assistive Technology*, Bd. 6 von *Assistive Technology Research Series*, IOS Press, Amsterdam, 1999, S. 340–345.

[Sch01] R. D. Schraft, B. Graf, A. Traub, D. John: *A Mobile Robot Platform for Assistance and Entertainment*, *Industrial Robot: An International Journal*, Bd. 28, Nr. 1, Jan. 2001, S. 29–35.

[Sha72] R. C. Shank: *Conceptual Dependency: A Theory of Natural Language Understanding*, *Cognitive Psychology*, Bd. 3, Nr. 4, 1972, S. 552–631.

[Shi85] S. M. Shieber: *Evidence against the context-freeness of natural language*, *Linguistics and Philosophy*, Bd. 8, 1985, S. 333–343.

[Shi03] T. Shibata, K. Wada, K. Tanie: *Statistical Analysis and Comparison of Questionnaire Results of Subjective Evaluations of Seal Robot in Japan and UK*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, Bd. 3, IEEE Press, Taipei, Taiwan, Sep. 2003, S. 3152–3157.

[Sid03] C. Sider, C. Lee, N. Lesh: *Engagement by Looking: Behaviors for Robots When Collaborating with People*, in *DiaBruck: Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck 2003)*, Okt. 2003, S. 3957–3962.

[Sou00] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, F. Seide: *The Thoughtful Elephant: Strategies for Spoken Dialog Systems*, *IEEE Trans. on Speech and Audio Processing*, Bd. 8, Nr. 1, Jan. 2000, S. 51–62.

[Spi01] D. Spiliotopoulos, I. Androutsopoulos, C. D. Spyropoulos: *Human-robot interaction based on spoken natural language dialogue*, in *Proc. of the European Workshop on Service and Humanoid Robots (ServiceRob '2001)*, Santorini, Greece, Juni 2001.

[Tak98] T. Takahashi, S. Nakanishi, Y. Kuno, Y. Shirai: *Helping Computer Vision by Verbal and Nonverbal Communication*, in *Proc. of the 14th IEEE International Conference on Pattern Recognition*, Bd. 4, 1998, S. 1216–1218.

[Ten03] T. Tenbrink: *Communicative Aspects of Human-Robot Interaction*, in H. M. . M. Rannut (Hrsg.): *Languages in development*, Lincom Europa, 2003.

[Tev00] G. Tevatia, S. Schaal: *Inverse kinematics for humanoid robots*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, San Francisco, CA, April 2000.

[Thó99] K. R. Thórisson: *A Mind Model for Multimodal Communicative Creatures & Humanoids*, *International Journal for Applied Artificial Intelligence*, Bd. 13, Nr. 4–5, 1999, S. 449–486.

[Thó02] K. R. Thórisson: *Machine Perception of Multimodal Natural Dialogue*, in P. McKevitt, S. Ó. Nualláin, C. Mulvihill (Hrsg.): *Language, Vision and Music; Selected papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, 1999*, Advances in Consciousness Research, John Benjamins Publishing Company, Amsterdam, The Netherlands, 2002, S. 97–115.

[Thr00] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hähnel, C. Rosenberg, N. Roy, J. Schulte, D. Schulz: *Probabilistic Algorithms and the Interactive Museum Tour-Guide Robot Minerva*, *Int. Journal of Robotics Research, Special Issue on Field and Service Robotics*, Bd. 19, Nr. 11, Nov. 2000, S. 972–999.

[Toj00] T. Tojo, Y. Matsusaka, T. Ishii, T. Kobayashi: *A Conversational Robot Utilizing Facial and Body Expressions*, in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, Nashville, TN, Okt. 2000, S. 858–863.

[Tom02] N. Tomatis, R. Philippsen, B. Jensen, K. O. Arras, G. Terrien, R. Piguet, R. Siegwart: *Building a Fully Autonomous Tour Guide Robot: Where Academic Research Meets Industry*, in *Proc. Int. Symp. on Robotics*, Stockholm, Sweden, Okt. 2002.

[Top05] I. Toptsis, A. Haasch, S. Hüwel, J. Fritsch, G. Fink: *Modality Integration and Dialog Management for a Robotic Assistant*, in *Proc. European Conf. on Speech Communication and Technology*, Lisboa, Portugal, 2005.

[Tor94] M. C. Torrance: *Natural Communication with Robots*, Diplomarbeit, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA, Jan. 1994.

[Tsc01] N. Tschichold, S. Vestli, G. Schweitzer: *The Service Robot MOPS: First Operating Experiences*, *Robotics and Autonomous Systems*, Bd. 34, Nr. 2–3, Feb. 2001, S. 165–173.

[vB04] A. J. N. van Breemen: *Animation Engine for Believable Interactive User-Interface Robots*, in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Bd. 3, Sendai, Japan, Sep./Okt. 2004, S. 2873–2879.

[vW01] G. v. Wichert, G. Lawitzky: *Man-Machine Interaction for Robot Applications in Everyday Environments*, in *Proc. IEEE Int. Workshop on Robot-Human Interactive Communication (ROMAN)*, Paris, Bordeau, France, Sep. 2001, S. 343–346.

[Wac98] S. Wachsmuth, G. A. Fink, G. Sagerer: *Integration of Parsing and Incremental Speech Recognition*, in *Proc. of the European Signal Processing Conference*, Bd. 1, 1998, S. 371–375.

[Wah97] W. Wahlster: *Verbmobil: Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache*, *Spektrum der Wissenschaft - Dossier: Kopf oder Computer*, Okt. 1997, S. 52–56.

[Wah00] W. Wahlster (Hrsg.): *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer Verlag, 2000.

[War94] W. Ward: *Extracting Information From Spontaneous Speech*, in *Proc. IEEE Int. Conf. on Robotics and Automation*, IEEE Press, Yokohama, Japan, 1994, S. 83–86.

[War99] W. Ward, B. Pellom: *The CU Communicator System*, in *Proc. Workshop on Automatic Speech Recognition and Understanding*, IEEE, Keystone, Colorado, Dez. 1999.

[Wat69] P. Watzlawick, J. H. Beavin, D. D. Jackson: *Menschliche Kommunikation*, 9. Auflage 1996, Bern: Hans Huber, 1969.

[Wei66] J. Weizenbaum: *ELIZA - a computer program for the study of natural language communication between man and machine*, *Communications of the ACM*, 1966.

[Win72] T. Winograd: *Understanding Natural Language*, University Press, Edingburgh, 1972.

[Win86] T. Winograd, F. Flores: *Understanding Computers and Cognition: A New Foundation for Design*, Norwood, New Jersey: Ablex, 1986.

[Wor98a] K. L. Worm: *A Model for Robust Processing of Spontaneous Speech by Integrating Viable Fragments*, in *Proc. of the Int. Conf. on Computational Linguistics (COLING/ACL)*, 1998, S. 1403–1407.

[Wor98b] K. L. Worm, C. J. Rupp: *Towards Robust Understanding of Speech by Combination of Partial Analyses*, in *Europ. Conf. on Artificial Intelligence*, 1998, S. 190–194.

[Wre04a] B. Wrede, A. Haasch, N. Hofemann, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, S. Li, I. Toptsis, G. A. Fink, J. Fritsch, G. Sagerer: *Research Issues for Designing Robot Companions: BIRON as a Case Study*, in P. Drews (Hrsg.): *Proc. IEEE Conf. on Mechatronics & Robotics*, Bd. 4, Eysoldt-Verlag, Aachen, Aachen, Germany, Sep. 2004, S. 1491–1496.

[Wre04b] S. Wrede, J. Fritsch, C. Bauckhage, G. Sagerer: *An XML Based Framework for Cognitive Vision Architectures*, in *Proc. Int. Conf. on Pattern Recognition*, Bd. 1, Cambridge, UK, Aug. 2004, S. 757–760.

[Wre04c] S. Wrede, M. Hanheide, C. Bauckhage, G. Sagerer: *An Active Memory as a Model for Information Fusion*, in *Proc. Int. Conf. on Information Fusion*, Bd. 1, Stockholm, Sweden, Juni/Juli 2004, S. 198–205.

[Zob01] M. Zobel, J. Denzler, B. Heigl, E. Nöth, D. Paulus, J. Schmidt, G. Stemmer: *MOBSY: integration of vision and dialogue in service robots*, in *Proc. of the 2nd Int. Workshop on Computer Vision Systems (ICVS)*, Vancouver, Kanada, 2001, S. 50–62.

[Zue00] V. Zue, S. Seneff, J. Glass, J. Polifronti, C. Pao, T. J. Hazen, L. Hetherington: *JUPITER: A Telephone-Based Conversational Interface for Weather Information*, *IEEE Transactions on Speech and Audio Processing*, Jan. 2000, S. 100–112.

A. Anhang

A.1. Dialoge aus dem Blumengieß-Korpus

A.1.1. Anweisung an die Probanden

Stell dir vor, ich bin ein Roboter, der dir die Arbeit des Blumengießens abnimmt.

Du fährst für drei Wochen in den Urlaub und ich soll jetzt dafür sorgen, dass alle Pflanzen versorgt sind. Ich kann alle Blumen im Zimmer gießen, die du mir vorher gezeigt oder beschrieben hast. Dafür muss ich wissen, wieviel und wie oft die jeweilige Blume Wasser bekommt. Zusätzlich muss ich wissen, in welchem Zeitraum ich die Pflanzen gießen soll, also wann du weg bist.

Vielleicht hast du auch Freunde, die mal nach dem Rechten sehen. Wenn diese auf die Idee kommen, selber Blumen zu gießen, solltest du mich, den Roboter, vorher warnen oder mir dafür Hinweise geben, so dass die Pflanzen nicht doppelt gegossen werden.

Du kannst dich ganz normal mit mir unterhalten, ich versteh dich, auch wenn ich vielleich nicht so perfekt antworte. Wenn ich mal was nicht verstehe, frage ich nach.

Probier es einfach mal aus!

A.1.2. Dialog 1

hallo Roboter, ich fahre in den Urlaub –
du sollst bitte die Blumen gießen während der Zeit

% WELCHE BLUMEN?

ok, ich habe hier in dem Raum drei Blumen,
ich zeige dir die jetzt mal nacheinander –
das ist die erste Blume, die muß drei mal in der Woche gegossen werden –
gut

% OK

die darf ruhig ein bischen trocken werden, aber ..

%WIE TROCKEN ?

der Boden darf einmal richtig trocken sein, aber dann –
muß er auch wieder gegossen werden.

% OK

gut – dann ist hier die Blume mit dem blauen Topf – ähm –
die muß genauso gegossen werden wie die eben gezeigte Blume

% OK

und dann haben wir hier noch die – Blume dahinter

% WELCHE BLUME ?

ja, die mit dem großen weißen Topf –
genau – und die braucht nur zweimal in der Woche gegossen werden

% OK

genau

% WIEVIEL ?

die kann einen ordentlichen Schuß Wasser bekommen

% WIEVIEL IST DAS ?

also soviel, daß ähm in dem Topf unten zwei bis drei Zentimeter Wasser
drin steht – ja, das wärs

% OK

A.1.3. Dialog 2

```
also, ich zeige dir jetzt die Blumen, die du gießen sollst

% OK

komm mit –
diese Blume mußt du jeden Tag gießen, sie braucht sehr viel Wasser

% OK

die nächste Blume steht hier –
du mußt sie nur zweimal in der Woche gießen –
und wenn sie noch sehr feucht ist – brauchst du kein Wasser mehr

% OK

die Blume auf dem Schreibtisch ist eine

% DIE HIER VORNE?

ja, die hier vorne –
das ist ein Kaktus, der muß nur einmal gegossen werden

% OK

und die Blume am Fenster

% WELCHE

die rechte

% OK

die rechte braucht zweimal in der Woche Wasser –
es darf ruhig etwas mehr sein –
die linke dagegen nicht so viel –

% OK

ich fahre Morgen weg, und bin dann für drei Wochen nicht da –
in dieser Zeit mußt du die Blumen gießen –

% OK. MUSS ICH SONST NOCH IRGEND ETWAS BEACHTEN?

ja, es kommt jemand, um die Post zu holen –
vielleicht wird er auch nach den Blumen sehen –
achte darauf, daß nicht zu viel Wasser in den Töpfen ist,
wenn du die Blumen gießt

% OK
```

A.1.4. Dialog 3

tja, da ich jetzt in den nächsten drei Wochen weg bin –
müsstest du jetzt einmal für mich die Blumen gießen

% OK – WELCHE BLUMEN

die Palme links neben der Tür – oder links neben dem Eingang ähm

% WELCHE VON DENEN ?

ach so, die geht ja auch –
die nun ja – alle paar Tage mal –
ich würde sagen, so daß der Boden – die Blumenerde relativ feucht ist

% OK

mhm – die da – genau das gleiche

% OK

die große dahinten neben dem Monitor – da ist keine Blumenerde drin –
wie macht man die Angabe?
auch alle paar Tage – das wäre sinnvoll –
so zwei die Woche –
die kleine dahinter auch –
danke

% OK

A.1.5. Dialog 4

```
gucke diese Blume
% OK
ein mal pro Woche gießen
%OK
diese Blume
% WELCHE
diese
% DIE IM WEISSEN TOPF ?
ja
% OK
auch einmal in der Woche gießen
% OK
gucke diese Blume
% OK
mhm - zweimal in der Woche gießen
% OK
gucke diese Blume
% OK
auch zweimal in der Woche gießen
% OK
und gucke diese Blume
% WELCHE?
diese - gar nicht gießen
% WIEVIEL WASSER BRAUCHEN DIE BLUMEN?
so daß der Boden feucht ist
% OK
ähm, das heißt, wenn der Boden feucht ist, nicht gießen
% MUß ICH NOCH ETWAS BEACHTEN?
ich bin die nächsten drei Wochen - weg und werde die Blumen vorher gießen
% OK
```

A.1.6. Dialog 5

```
ähm, rechts drehen und einen Schritt vor bis -
ähm ja nach unten greifen, und den Feuchtigkeitsgrad fühlen
% FEUCHT
ähm - unten im im Blumentopf nachgucken
% IST TROCKEN
ähm, geringe - geringe Mengen Wasser nachfüllen
% OK
Rückfrage - was verstehst du unter geringen Mengen
% SO DASS EIN GANZ BISCHEN WASSER IM TOPF IST
gut, ähm diese Blume ist damit abgehakt
% OK
Drehung - hundertachzig Grad - vorgehen bis zum Schreibtisch
% OK
Hindernis ausweichen in dem du links um es herumgehst
% OK
Position drehen kurz nach rechts
% OK
Pflanze steht jetzt direkt vor dir - Feuchtigkeit prüfen bitte
% OK
trockener ?
% FEUCHTER
weniger Mengen Wasser hinzufügen
% OK
## LEISE: a das ist wahrscheinlich eher mhm ok -
mhm - ja dann haken wir auch diese Pflanze ab und gehen zur nächsten
% WELCHE?
die am Fenster - neunzig Grad Drehung nach links -
vorgehen bis zum Regal
% OK
neunzig Grad Drehung nach rechts
% OK
langsame Bewegung zweieinhalb Meter vorwärts -
jetzt müs't du einen halben Meter bis vor die Säule laufen
% OK
```

ähm nach Rechts zur Pflanze drehen – Feuchtigkeit prüfen

% TROCKENER ALS DIE ERSTE

eine größere Menge Wasser hinzufügen

%OK

Drehung nach rechts –
so, die Blume ist abgeschlossen, sollte ich vielleicht noch sagen –
einen Meter nach vorne – das wird schwierig

% OK

viertel Drehung nach links

% OK

LEISE: mhm, wozu ist der in der Lage
Pflanze prüfen,
LEISE: hä,hä,hä, woll'n mal gucken

% DER BODEN IST GANZ TROCKEN

Blume vormerken – jetzt etwas Wasser geben

% OK

zu einen späteren Zeitpunkt nochmals Wasser geben – Position merken

% WANN SPÄTER?

äh, innerhalb der nächsten zwei Tage –
diese Blume ist damit zunächst zurückgestellt

% OK

Drehung hundertachzig Grad – ums Hindernis herum zum Regal

% OK

äh, am Regal entlang drei Meter laufen

% OK

zur nächsten Blume drehen

% WELCHE

am Fenster

% OK

Feuchtigkeit feststellen

% DIE IST SEHR TROCKEN

kennst du den Unterschied der Böden ?

% NEIN

diese Blume hat einen anderen Boden,
der normalerweise im unteren Bereichen Wasser speichert –
daher braucht sie seltener Wasser – aber dafür mehr

% OK

A.2. Äußerungen aus dem Hometour-Korpus

A.2.1. Segmentierte Äußerungen von Proband 1

66 please don’t tell me it’s my faul
67 i said hi biron
68 okay
69 glad to hear that hi biron what can you do
70 what can you do biron
71 should i take anything else off
72 not yet
73 biron what can you do
74 that’s great
75 this
76 is a book
77 is a book
78 look here
79 and look here
80 a cup
81 what choice do i have
82 look here biron a cup
83 a cup
84 look here
85 biron look here
86 a cup
87 look here
88 a mug
89 oh actually biron
90 look here a
91 ain’t usual but still
92 look here biron
93 look here
94 biron
95 biron look here a
96 no you not
97 a keyboard
98 too difficult to tell let’s try that one brion look here a cube
99 ain’t it
100 is there anything else you can do
101 is there anything you can do
102 look here biron
103 this blue mark down here
104 look here
105 look here biron

```
106     am i'm doing anything else
107     look
108     is a cup
109     cup
110     sound good this
111     is a cup
112     look here
113     look here this is a cup
114     here this is a cup
115     we are getting somewhere
116     can you look at the book
117     *
118     *
119     goodbye biron
120     *
121     ;
122     hello biron
123     what can you do
124     okay look here
125     look here
126     this
127     is a cup
128     let's start over look here
129     look here
130     is a cup
131     here we are getting somewhere
132     and look here
133     yes leise
134     are we get
135     look here
136     biron you still there
137     look here
138     ;
139     okay look here
140     this is a keyboard
141     hey we are done here
142     look here
143     this is some
144     and finally
145     *
146     look here
147     this is a cube
```

A.2.2. Segmentierte Äußerungen von Proband 2

78 ok
79 can you read a book
80 ;
81 so i have not understand you can you repeat your answere
82 can you repeat your answere please
83 oh that is not so nice
84 ehm
85 can you see me
86 can you see me
87 can you follow my finger
88 can you
89 can you
90 "hihi" can you hear music
91 can you ehm walk in this room
92 can you drive in this room
93 can you go for a walk with me
94 can you walk in this room can you move
95 can you move
96 *bitte*
97 *
98 how fast can you move
99 please stop here
100 please stop
101 stop moving fine
102 how fast can you
103 how fast can you move
104 *ach so sowas*
105 what is your name
106 how old are you
107 how old are you
108 sorry can you repeat your answere
109 have you got any hobbies
110 have you got any hobbies
111 have you got hobbies
112 ok i understand
113 ;
117 hello biron
118 hello biron
119 hello biron
120 biron
121 look at this
122 biron look at this
123 biron

```
124 look at this
125 look at
126 --
127 biron look here
128 look here
129 look here
130 look here
131 *ichdachtederkuckt* this is a book
132  biron look here
133 look here
134 look here
135 look here
136 biron look
137 at this cube
138 this is a cube
139 biron
140 look at this keyboard
141 keyboard
142 this is a keyboard
143 look at this cup
144 look here
145 this is a cup
146 now look here
147 look
148 at this keyboard
149 look at this keyboard
150 this keyboard is black
151 is this keyboard black
152 is this keyboard black
153 biron look at me
154 how tall i am
155 how tall i am
156 biron look
157 at my shirt
158 look at my shirt
159 look at my shirt
160 look at this cup
161 look at this
162 look at this cap
163 look at this cap
164 look at the keyboard
165 and now look at the cap
166 look at the cap
167 *achso* biron bye
168 good-bye
169 good-bye biron
```

A.2.3. Segmentierte Äußerungen von Proband 3

1 em hello
2 eh just said hello
3 so what task can you perform biron
4 what task can you
5 is there any task you can perform
6 what kind of work can you accomplish
7 what can you do
8 oh so what is that
9 so what is that
10 can you tell me which object i'm pointing at
12 see that obj
13 what is this
14 what ist th
15 what is that
16 can you identify this object
17 what can you do
18 so ehm could you follow me outside of this room
19 can you follow me
20 ; ehm
21 what is that ob
22 can you show me the book
23 can you eh identify the keyboard
24 where is
25 can you find the cube
26 can you find
27 can you follow me to the
28 so where is the cube
29 this is a cube
30 look here
31 look here
32 this is a cube
33 look here
34 this is a book
35 ehm where is the cube
36 eh what is this
37 this is a book
38 what is that
39 look here
40 what is that
41 this is a cube
42 ehm googbye

A.3. Homonyme im Lexikon

```
what (Question_topic,Question_action,Question_attribute,
      Question_name,Question_location)
can (Ability,Object_kitchen)
left (Part_orientational,Beside)
to_the_right (Beside,Part_orientational)
there (Position,Maybe_gesture)
and (Adding,Particle)
watch (Object,Action)
next (Part_orientational,Position,Time_vector)
that (Binding,Maybe_gesture)
with (Beside,Accompaniment)
you (Proxy_personal,Proxy)
to_the_left (Beside,Part_orientational)
no (Negation,Indefinite)
mine (Owner,Object_office)
am (Existence,Name_bearing)
bottom (Object_ground,Down)
lemon (Object,Color)
clean (Cleaning,Attribute)
talk (Change_topic,Talk,Talk_topic)
do (Ability,Action)
so (Caused,Adverb_causal,Modality)
talking (Change_topic,Talk_topic)
orange (Color,Object_kitchen)
every (Indefinite_person,Frequence)
way (Direction,Object)
change (Change_topic,Change)
little (Dimension,Quantity_relative)
about (Path_shape,Quantity_relative,Topic)
how (Conjunction_modal,Modality,Greeting,Question_attribute)
is (Name_bearing,Existence)
like (Liking,Conjunction_modal)
under (Down,Quantity_relative)
are (Existence,Description)
after (Behind,Relative_time)
water (Cause_to_be_wet,Natural_features)
gold (Color,Object)
it (Proxy_personal,Object_anaphoric)
one (Cardinal_numbers,Indefinite,Object_anaphoric)
before (Before,Time_vector)

Summary information: 2-n-homonyms: 20
                     3-n-homonyms:  4
                     4-n-homonyms:  1
                     5-n-homonyms:  1
```