Dynamische Themenerkennung in situierter Mensch-Roboter-Kommunikation

Jan Frederik Maas

April 2007

Schriftliche Arbeit zum Erwerb des Grades „Doktor der Ingenieurwissenschaften“
Technische Fakultät
Universität Bielefeld

Gutachter:
Dr.-Ing. Britta Wrede
Prof. Dr. Alexander Mehler

Inhaltsverzeichnis

Danksagung:

Ein herzliches „Danke!“ an alle, die diese Arbeit ermöglicht haben – auf wissenschaftlicher Seite an Britta Wrede, Alexander Mehler, Gerhard Sagerer und „Max“ Sichelschmidt, an das Graduiertenkolleg „Aufgabenorientierte Kommunikation“, an die Mitglieder der Arbeitsgruppe „Angewandte Informatik“– insbesondere Axel Haasch und Joachim Schmidt, COGNIRON, die DFG so wie die Universität Bielefeld.

Auf privater Seite möchte ich ebenfalls herzlich meinen Eltern Renate und Gerhard Maas danken so wie Silke Fürhoff für freundschaftliche Ratschläge und für die große Menge an guter Laune, die mir während der vergangenen Jahre geschenkt wurde.

Erklärung:

In der deutschen Schriftsprache hat sich in der Vergangenheit die Verwendung des Maskulinums als nützlich dahingehend herausgestellt, dass es für „männliche“ und „weibliche“ Nomina als neutraler Oberbegriff verwendet werden darf. Aus diesem Grund möchte ich in dieser Arbeit auf komplizierte Konstruktionen wie „Studenten/Innen“ verzichten. Natürlich sind stets sowohl weibliche als auch männliche Personen gemeint, wenn sich anderes nicht direkt aus dem Zusammenhang erschließen lässt (cf. „die männlichen Versuchspersonen“).

Gedruckt auf alterungsbeständigem Papier nach ISO 9706.

1 Einleitung

Nur wenige Bereiche aktueller Forschung haben in der Vergangenheit so viel öffentliches Interesse gepaart mit Emotionalität hervorgerufen, wie die Entwicklung von Robotersystemen, die menschenähnlich zu kommunizieren in der Lage sind. Ein Indiz für diese Behauptung ist die Omnipräsenz von Robotern in *science fiction*-Romanen und -Serien. Unabhängig davon, ob man die künstliche Maria aus „Metropolis“ von Fritz Lang aus dem Jahre 1927 betrachtet, R2D2 aus „Star Wars“, Tachikoma aus „Ghost in the Shell“ oder die aktuelle Neuauflage der Serie „Battlestar Galactica“, in der humanoide Roboter existieren, die nur mit extremem technischen Aufwand von Menschen zu unterscheiden sind – der Traum von einem technischen „Lebewesen“ scheint noch lange nicht ausgeträumt zu sein. Interessant ist in diesem Zusammenhang auch, dass der Begriff „Roboter“ Anfang des 20. Jahrhunderts von Josef und Karel Čapek geprägt wurde – im Rahmen der *science fiction*, weniger der Wissenschaft.

Abbildung 1.1: Kunststoffnachbildung von Tachikoma aus „Ghost in the Shell“

Den mythischen Vorläufer des Roboters stellt u.a. der Golem dar, dessen Mythos sich bis auf das 12. Jahrhundert zurückverfolgen lässt. Golems sind in Analogie zum jüdisch-biblischen Schöpfungsakt[1] aus Lehm geschaffene und magisch bewegte Humanoide, die wortgetreu den Befehlen ihres Erschaffers folgen. Aufgrund dieses Umstands können sie sowohl sehr hilfreiche als auch unheimliche Züge (im Kontext einer falschen „Programmierung“) annehmen.

Sowohl Golems als auch Roboter sind in der *science fiction* stets von Menschen (oder anderen intelligenten Wesen) geschaffen. Sie können unheimlich sein, wie der Golem, sich gegen ihre Schöpfer wenden, wie die Zylonen aus „Battlestar Galactica“, freundlich bis niedlich sein wie R2D2 aus „Star Wars“ – emotional neutral besetzt

[1] jüdisch-biblischer Mythos nach Genesis, 2-7

sind sie aber in den seltensten Fällen. Auffällig ist auch, dass in der Belletristik die Roboter sehr oft zumindest im Ansatz menschenähnliche Züge besitzen. Im Gegensatz dazu sind intelligente, technische Entitäten – die meist mittels natürlicher Sprache kommunizieren – oft ohne unabhängige Verkörperung anzutreffen. Ein Beispiel hierfür ist der Schiffscomputer HAL aus Space Odyssey 2001 von Stanley Kubrick aus dem Jahre 1968. (Trotzdem kann HAL durchaus als Roboter gesehen werden, dessen Verkörperung das Raumschiff selbst ist).

Verlässt man das Gebiet der utopischen Literatur, so findet man Roboter vor allem in den Bereichen der industriellen Fertigung. Hier ist von einem emotionalen Bezug der Benutzer zu diesen Maschinen wenig festzustellen – Aspekte wie Funktionalität, Zuverlässigkeit etc. stehen weitaus mehr im Vordergrund. Trotzdem lässt sich in den letzten Jahren eine Tendenz beobachten, Roboter im übertragenen Sinne „aus den Fertigungshallen hervorzuholen“ und ihnen einen Einzug als Helfer, Spielzeuge oder sogar Pflegeroboter in Haushalte zu erlauben. Mit den industriell genutzten Robotern haben diese „Wesen“ oft nur noch wenig gemeinsam, das einzige verbindende Element ist die für Roboter per Definition notwendige Fähigkeit zur quasi-autonomen Bewegung.

1.1 Roboter im Haushalt

Im Gegensatz zu einem Industrieroboter wird an einen Haushaltsroboter der Anspruch gestellt, dass der Besitzer eine positive emotionale Einstellung zu dem Gerät haben sollte. Diese ist oftmals nicht nur wünschenswert, damit die Besitzer zufriedener sind und das Gerät einen besseren Absatz findet, bisweilen ist sie sogar zur Erfüllung der jeweiligen Aufgabe notwendig – z.B. bei Spielzeugrobotern. Grundsätzlich sind also zwei verschiedene, parallele Entwicklungen zu erwarten: Zum einen werden in der Zukunft bestimmte Roboter in Haushalten ihren Dienst so unauffällig wie möglich verrichten, damit keine Störung des Alltagslebens eintritt. Hochspezialisierte Roboter-Staubsauger, wie sie heutzutage schon erhältlich sind (s.u.), sind ein Beispiel für diese Art von Maschinen: Der Roboter erzeugt nur eine minimale negativ-emotionale Disposition durch Lärm, nötige Aufmerksamkeit etc. und eine stark positive durch die Erfüllung von einer (lästigen) Aufgabe.

Auf der anderen Seite wird es Robotersysteme geben, die aus bestimmten Gründen zwangsläufig in den Aufmerksamkeitsfokus ihrer Besitzer treten müssen: Ein gegenwärtiges Beispiel sind Roboterspielzeuge, Roboterdiener und verkörperte Agenten, die ein *interface* zur Haustechnologie und zum Terminkalender darstellen, wie z.B. die Roboterkatze iCat von Philips (vgl. Abbildung 1.2 auf der nächsten Seite). Die letzten beiden Gruppen – Roboterdiener und verkörperte Agenten – benötigen einfach aufgrund der Komplexität ihrer Aufgaben die Kommunikation mit dem Benutzer. Im Gegensatz zu einem hochspezialisierten Staubsauber ist ihre Aufgabe eben nicht eindeutig und durch das Wissen lösbar, welches sie werksseitig erhalten haben.

Trotz der Reduktion der Betrachtung auf Roboter, die dazu gedacht sind, in Haushalten ihre Arbeitsumgebung zu finden, bleibt eine außerordentlich große Vielfalt von schon existierenden Typen übrig. In grober Anlehnung an (Fong u. a., 2002) möchte ich hier zur Übersicht eine Klassifikation nach Anwendungsgebiet vorstellen:

Abbildung 1.2: Mensch-Maschine-Interfaceroboter iCat

Serviceroboter Serviceroboter sollen Aufgaben in Haushalten erfüllen, für die Mobilität ein wichtiges Kriterium ist. Bekannte, mittlerweile schon käuflich zu erwerbende Geräte dieser Art sind Rasenmäher oder die bereits erwähnten Staubsauger. Meist handelt es sich bei diesen Geräten um scheibenförmige, autonom umherfahrende Maschinen, die den unter ihnen befindlichen Boden reinigen bzw. das Gras mähen. Ein Beispiel ist Roomba von der Firma iRobot (vgl. Abbildung 1.3 auf der nächsten Seite). Serviceroboter sind im Allgemeinen nicht auf ein äußeres Erscheinungsbild festgelegt – im Wesentlichen existiert nur die Anforderung, dass der Roboter seiner Aufgabe nachgehen kann. Im Rahmen dieser Betrachtung wäre ein humanoider Roboter, der einen Rasenmäher schiebt, nicht zwangsläufig (bezüglich der Kernfunktionalität) ein besserer oder schlechterer Seviceroboter im Vergleich zu einem scheibenförmigen Gerät – nur komplizierter und teurer zu konstruieren und ggf. angenehmer für den Benutzer.

Therapeutische Roboter Therapeutische Roboter sollen medizinischen Zwecken dienen, so z.B. der besseren Unterstützung Kranker in Alltagsangelegenheiten. Ein Beispiel für einen solchen Roboter ist der in (Prassler u. a., 1999) dargestellte Roboterrollstuhl. Alternativ können therapeutische Roboter auch die Züge eines Spielzeugs annehmen. Das wohl bekannteste Beispiel für ein therapeutisches Roboterspielzeug ist die Robbe Paro, die speziell in Altenheimen und in Kinderkrankenhäusern mit Erfolg zum Einsatz gekommen ist (vgl. Abbildung 1.4 auf Seite 5, (Wada u. a., 2002)).

Unterhaltungsroboter/Spielzeuge Natürlich existiert eine große Anzahl von Spielzeugrobotern, deren Unterhaltungswert Selbstzweck ist. Das wohl bekannteste Beispiel für einen Spielzeugroboter ist der Roboterhund AIBO, der von der Firma Sony

Abbildung 1.3: Staubsauger Roomba von der Firma iRobot
(entnommen der URL: http://www.tomorrow.de/pc/hardware/roboter-staubsauger ?slide=8&interface=slide&ao_id=1994&id=1146&page=1 am 22. 11. 2006)

entwickelt wurde (vgl. Abbildung 1.5 auf der nächsten Seite). AIBO beherrscht verhältnismäßig komplexe Handlungen wie z.B. das Aufnehmen eines Plastikknochens oder das eigenständige Andocken an die Ladestation. Ein weiteres Merkmal von AIBO ist seine Fähigkeit, einfache gesprochene Sprache zu verstehen. Diese Fähigkeit orientiert sich an den Ein-Wort-Befehlen, die echten Hunden gegeben werden können, und ist somit vom Verständnis komplexer Befehle weit entfernt. Trotzdem kann davon ausgegangen werden, dass durch diese Fähigkeit eine viel stärkere Identifikation des Besitzers mit dem Roboter erzielt werden kann, als es bei einem stummen Modell der Fall wäre.

Ich möchte in dem nächsten Abschnitt stärker auf diesen wichtigen Aspekt – dem Nutzen von verbaler Kommunikation mit Robotersystemen – eingehen.

1.2 Roboter und Sprache

Wie weiter oben geschildert müssen Serviceroboter bei vielen Aufgaben in der Lage sein, mit ihren Besitzern zu kommunizieren. Kommunikation kann durch das Eintippen komplexer Steuerbefehle in eine an den Roboter angeschlossene Tastatur erfolgen, aber ein solches Vorgehen führt aufgrund der Nichttrivialität schnell zu einer Frustration – also einer negativen emotionalen Disposition – bei dem Benutzer. Diese Nichttrivialität kommt aufgrund von drei Faktoren zustande:

1. Die Lösung des Problems muss von Menschen vorgegeben werden.
2. Tastatureingaben sind (in bestimmten Situationen) unhandlich.
3. Der Mensch muss einen nicht unerheblichen Lernaufwand aufbringen, um die Steuercodes zu erlernen.

Der erste Punkt wird von Jörg Pflüger in (Pflüger, 2004) untersucht. Pflüger unterscheidet chronologisch drei Stufen von Interaktionsidealen mit Computersystemen: **Konversation**, **Manipulation** und **Delegation**.

Abbildung 1.4: Therapeutischer Roboter Paro
(entnommen der URL: http://www.fathom.com/feature/122598/3590_paro_LG.html am 22. 11. 2006)

Im Kontext der frühen Entwicklung von Computersystemen kam der Gedanke auf, dass Roboter als kommunikativ nahezu gleichwertige Partner in der Lage seien, den Menschen direkt – im Rahmen einer **Konversation** – bei der Lösung eines Problems zu helfen. Dieses Bild scheiterte schnell an den realen Möglichkeiten von Computersystemen und wurde aufgrund der Entwicklung von Desktoprechnern durch das Bild der direkten **Manipulation** ersetzt: Ein Programmierer hat die wesentlichen Problemlösungsstrategien schon vorgegeben, der Benutzer manipuliert auf dem Bildschirm Objekte, die seinem Anwendungsfall entsprechen.

Durch die zunehmende Verbreitung von Agenten und Avataren scheinen wir an dem Anfang eines neuen Paradigmas zu stehen: Dem Bild der **Delegation**. Der Computer ist intelligenter geworden und kann bestimmte Aufgaben halbintelligent lösen. Der Unterschied zwischen manipulativ und delegativ gesteuerten Robotersystemen kann anhand des folgenden Beispiels skizziert werden: Im Fall der Manipulation bewegt der Benutzer mit Hilfe detaillierter Befehle einen Roboterarm, bis dieser ein Objekt aufgehoben hat. Im Fall der Delegation reicht es, den Roboter über die Aufgabe zu

Abbildung 1.5: Roboterspielzeug AIBO ERS-7M2 von Sony

informieren („hebe bitte das Objekt auf“).[2]

Der zweite oben angeführte Punkt – die Unhandlichkeit der Tastatur oder vergleichbarer Eingabegeräte – sollte im Idealfall durch einen Verzicht auf solche Geräte angegangen werden. An dieser Stelle stellt sich wiederum menschliche Sprache als das aus Benutzersicht ideale Kommunikationsmedium dar.

Der letzte Punkt – der Lernaufwand, der Voraussetzung für eine erfolgreiche Kommunikation über Steuerbefehle ist – wird im Falle eines Robotersystems, welchem Aufgaben delegiert werden können und mit dem man über natürliche Sprache kommunizieren kann, ebenfalls weitgehend aufgelöst. Zusammenfassend kann also festgestellt werden, dass der „ideale“ Haushaltsroboter – sofern er verstärkte Interaktion mit seinem Besitzer zur Bewältigung seiner Aufgaben benötigt – ein natürlichsprachlich kommunizierendes System sein sollte, das in gewissem Maße zur eigenständigen Lösung ihm delegierter Aufgaben in der Lage ist.

1.3 Zielsetzung

Natürlichsprachlich kommunizierende Robotersysteme, die intelligent in Haushaltsumgebungen ihnen delegierte Aufgaben erledigen, sind sowohl ein erstrebenswertes Ziel als auch ein interessanter Forschungsgegenstand. Zu beachten ist, dass es in der Fähigkeit, natürlichsprachlich zu kommunizieren, große Qualitätsunterschiede geben kann: AIBO ist in der Lage, natürliche Sprache zu verstehen (und vorgegebene Sätze zu produzieren), aber seine Fähigkeiten sind mit denen eines menschlichen Haushälters oder Butlers nicht vergleichbar. Umgangssprachlich kommunizierende Dialogsysteme, die in der Lage sind, natürlichsprachliche Äußerungen zu verstehen und zu produzieren, sind in den letzten Jahren in großer Zahl erstellt worden, trotzdem kommen mit jedem Teilaspekt natürlichsprachlicher Kommunikation, zu dem ein neues Dialogsystem in der Lage ist, noch nicht behandelte Aspekte in den Fokus der Forschung. Aus diesem Grund sind Dialogsysteme gerade für die Linguistik ein fruchtbares Forschungsfeld, in dem entwickelte Theorien mit verhältnismäßig geringem Aufwand getestet werden können. Der Schritt von einem nicht-verkörperten Dialogsystem zu einem virtuellen Avatar oder sogar einem Roboter eröffnet aber noch einmal eine neue Dimension von Forschungsfragen, da auf diese Weise wichtige zusätzliche Modalitäten der Kommunikation wie Mimik oder Gestik, aber auch die Situiertheit – also die Eingebundenheit der Kommunikationspartner in eine Kommunikationssitation – relevant werden.

Diese Arbeit macht es sich zur Aufgabe, einen neuen Aspekt von situierter Mensch-Roboter-Kommunikation zu untersuchen: Der Erkennung von Themen. Die Zielsetzung dieser Arbeit lässt sich somit verhältnismäßig einfach in einem Satz zusammenfassen: Im Rahmen dieser Arbeit sollte ein System zur dynamischen *online*-Themenerkennung in situierter, aufgabenorientierter und natürlichsprachlicher Kommunikation für einen Roboter entwickelt und evaluiert werden. Ein solches System kann die kommunikativen Fähigkeiten eines Roboters stark unterstützen: Abgesehen von der Disambiguierung von nur situativ interpretierbaren Äußerungen sind auch basalere Anwendungsfälle wie die Unterstützung der Spracherkennung denkbar. Auf

[2] Der Fall der Konversation ist im Rahmen dieses Beispiels schlecht darstellbar, ggf. würde der Roboter selbst vorschlagen, dass und wie das Objekt (in Koorperation mit dem Benutzer) aufzuheben sei.

diese Weise unterstützt ein Themenerkennungssystem nicht nur die Fähigkeit des Robotersystems, natürlichsprachlich zu kommunizieren, sondern auch die Fähigkeit der Umsetzung delegierter Aufgaben: Im Falle der (delegativen) Anweisung „kümmere dich bitte um das Geschirr“, die nur situativ-thematisch interpretierbar ist, wird dieser Aspekt deutlich. Sie kann je nach Kontext die Aufforderung, das Geschirr zu spülen, es wegzuräumen oder aufzudecken beinhalten. Eine eingehendere Diskussion der genannten Punkte findet sich in Abschnitt 2.2; der interessierte Leser sei an diese Stelle verwiesen.

Themenerkennung hat sich in den letzten Jahren als ein sehr aktives Forschungsfeld etabliert. Klassische Anwendungsszenarien für Themenerkennungssysteme sind z.B. die Sortierung von Artikeln aus Printmedien nach Themen, die so eine Filterung der Daten oder eine schnelle Suche in der Menge der Artikel erlauben. Ein Forschungsprojekt der Vereinigten Staaten – das „Topic Detection and Tracking“ (TDT)-Projekt (Allan, 2002b) – hat sich mit großem Erfolg an der thematischen Klassifikation von Radionachrichten versucht. Ein Anwendungsszenario für ein TDT-System wäre z.B. ein intelligentes Autoradio, welches nach bestimmten Themen sucht und selbstständig den Kanal wechseln kann. Natürlich existieren auch geheimdienstliche Anwendungen, wie z.B. die schnelle Sortierung von Abhörinformationen.

Themenerkennung im Bereich der *interaction robotics*, also Themenerkennung zur Unterstützung von kommunikativen Fähigkeiten von Robotern, ist ein neues Feld, zu dem ich im Rahmen dieser Arbeit beitragen möchte. Dabei habe ich weniger versucht, einen algorithmisch perfektionierten Ansatz zu entwickeln, als vielmehr die Grundlagen für zukünftige Forschung durch Erstellung eines funktionierenden Prototypensystems zu schaffen. Die Anforderungen an ein solches System sind in dem obigen Satz zusammengefasst. Ich möchte an dieser Stelle kurz auf die Begrifflichkeiten und die durch sie ausgedrückten Bedingungen eingehen:

Online ist ein Klassifikationsprozess, wenn er nahezu ohne Zeitverzögerung nach Eingang des zu klassifizierenden Signals ein Erkennungsresultat liefert. Ein Beispiel für einen *online*-Erkennungsprozess wäre ein Kind, dem man bunte Blätter zeigt und das sofort die jeweilige Farbe der Blätter benennt. Im Gegensatz dazu steht *offline*-Erkennung, bei der die Erkennung oder Klassifikation erst nach der Sammlung der Daten und möglicherweise mit großem Zeitaufwand stattfindet. Ein Beispiel für eine *offline*-Erkennung wäre eine OCR-Software, die erst längere Schriftstücke analysieren muss, um eine als Bild vorliegende Handschrift effektiv in Text umzuwandeln. *Offline*-Prozesse finden normalerweise in nicht-zeitkritischen Umgebungen statt. Für ein Roboter-Dialogsystem, welches mehr als nur statistischen Nutzen aus einer Themenerkennung ziehen will, muss eine Themenerkennung zwingend *online* implementiert sein.

Dynamische Themenerkennung ist nicht auf vorgegebene Themen oder eine vorbestimmte Anzahl von Themen festgelegt. In Umgebungen, in denen die Art und Anzahl der Themen stark von der situativen Gegebenheit abhängt, bringt eine dynamische Themenerkennung große Vorteile. Allerdings stellt Dynamizität eine weitere Herausforderung an Themenerkennungssysteme dar, da dynamische Systeme weitaus schwieriger und weniger effizient als statische Systeme zu verwirklichen sind.

Im Rahmen dieser Arbeit – der Entwicklung eines Themenerkennungssystems für einen Haushaltsroboter bzw. für in Haushalten tätige Roboter – ist es notwendig, das System dynamisch zu gestalten. Der Grund dafür liegt in der Unmöglichkeit, die

verschiedenen Themen, die in einem Haushalt vorkommen können, vorherzusagen. Definiert man z.B. Kaffeemaschinen (thematisch) als Teil der Küche, würde man Fälle fehlbehandeln, in denen diese ein Teil des Wohn- oder Schlafzimmers sind. Aus diesem Grund ist die Implementierung von Dynamizität eine Grundvoraussetzung dieser Arbeit.

Ein weiterer Aspekt von Dynamizität, nämlich der Wandel von Themen oder thematischen Zusammenhängen im zeitlichen Verlauf, wird in dieser Arbeit nur am Rande untersucht. Die verwendeten Algorithmen unterstützen aber auch diesen Aspekt.

Situierte Kommunikation ist bezogen auf die Situation – und damit auch auf Aspekte der Umgebung – der jeweiligen Kommunikation (Milde u. a., 1997). Ein Erkennungsmerkmal für situierte Kommunikation ist explizite Kommunikation über situationsbezogene Aspekte („sieh mal, der Tisch hier“); oftmals ist situierte Kommunikation ausschließlich innerhalb der Situation bzw. durch Einbeziehung von situativem Wissen verständlich. Ein Beispiel dafür sind Zeigegesten auf Objekte; eine Zeigegeste ist ohne das gezeigte Objekt unverständlich. Es ist evident, dass ein Haushaltsroboter über die Fähigkeit der situierten Kommunikation verfügen sollte.

Aufgabenorientierte Kommunikation dient der Erfüllung eines gemeinsamen Ziels der Kommunikatoren. Oft ist aufgabenorientierte Kommunikation situiert, d.h. das zu erreichende Ziel ist in der Situation zu verwirklichen. Mensch-Roboter-Kommunikation ist in den meisten Fällen aufgabenorientiert, auch wenn in der Zukunft sozial interagierende Roboter den Anteil an nicht-aufgabenorientierter Kommunikation vergrößern werden.

1.4 Überblick

Auch wenn Ziel und Nutzen dieser Arbeit in den vorangegangenen Sätzen im Wesentlichen beschrieben wurden, möchte ich trotzdem das folgende Kapitel 2 dazu verwenden, um auf Fragen, die diese sehr knappe Darstellung aufwirft, detaillierter einzugehen. Neben der Diskussion der Nutzenaspekte einer Themenerkennung für natürlichsprachlich kommunizierende Robotersysteme wird weiterhin versucht werden, eine Arbeitsdefinition des Begriffs „Thema“ zu finden.

Für Leser, die ausschließlich an der technischen Umsetzung des Systems und weniger an den linguistischen und kommunikationstheoretischen Grundlagen interessiert sind, empfiehlt es sich eventuell, mit den nachfolgenden Kapiteln fortzufahren. Selbige unterteilen sich wie folgt:

In Kapitel 3 wird die grundlegende Struktur des Themenerkennungssystems anhand von Vorarbeiten schrittweise entwickelt. Dazu werden in dem Kapitel die theoretischen Hintergründe dargestellt, wobei auf zur Anwendung gekommene Technologien wie semantische Räume eingegangen wird. Kapitel 4 beschreibt das BITT-Korpus, das als Evaluationsgrundlage im Rahmen dieser Arbeit erstellt wurde. Die Evaluationen auf diesem Korpus anhand eines *offline*-Prototyps des Themenerkennungssystems werden in Kapitel 5 beschrieben. Im Kapitel 6 wird die *online*-Implementierung des Systems auf dem Roboter BIRON geschildert, wobei auch die technische Grundlage – das Robotersystem – im Detail vorgestellt wird. Das Kapitel schließt mit der Darstellung einer experimentellen Evaluation der Themenerkennung im Rahmen einer Benutzerstudie. Anschließend folgt die Diskussion der Arbeit so wie ein Ausblick auf weitere Forschungen.

2 Motivation: Robot Companions und Themenerkennung

Diese Arbeit entstand in Assoziation mit dem COGNIRON[1]-Projekt der Europäischen Union. Das COGNIRON-Projekt ist ein internationaler, wissenschaftlicher Forschungsverbund. Das Forschungsziel des Projektes ist die Entwicklung von Robotern, die sich in menschlichen Alltagsumgebungen zurechtfinden und agieren können. Zentral für das Projekt ist dabei die Entwicklung von Robotern, die nicht nur als intelligente, spezialisierte Werkzeuge fungieren, wie es z.B. ein intelligenter Kühlschrank oder ein autonomer Staubsauer sein würde. Das Ziel von COGNIRON besteht in der Entwicklung von Robotern, die als intelligente *companions* arbeiten – ein solcher *companion* kann als eine Art technischer Begleiter und Helfer angesehen werden, der Menschen im Umgang mit ihrer sich stets ändernden Alltagswelt zur Seite steht. *Companions* müssen in der Lage sein, selbstständig ihre neue Umgebung zu erfassen, neue Fähigkeiten zu erlernen und dabei auch proaktiv mit den Menschen zu kommunizieren.

Im Rahmen des COGNIRON-Projektes wurden verschiedene Meilensteine definiert, anhand derer ein solcher Roboter entwickelt werden soll. Die Meilensteine werden durch standardisierte Szenarien bestimmt, in denen – teilweise aufeinander aufbauend – Basisfähigkeiten des Roboters zum Einsatz kommen. Ich möchte die Szenarien in den folgenden Absätzen kurz schildern.

Robot home tour Das „*robot home tour scenario*“ oder kurz *home tour*-Szenario stellte insbesondere für die Universität Bielefeld den Forschungsschwerpunkt dar. In einem *home tour*-Szenario muss ein Roboter mit Hilfe eines menschlichen Führers Aspekte seiner (Haushalts-)Umgebung erlernen. Diese Aspekte bestehen im Wesentlichen aus

- topologischen Informationen, also Informationen über Raumaufteilung, Bodenbeschaffenheit, Hindernisse, Höhenunterschiede etc.
- Art, Position und Relevanz für etwaige Handlungen der Objekte in der gegebenen Umgebung.

Der Unterschied besteht im Wesentlichen in der jeweiligen Relevanz von Aspekten der Umgebung für den Roboter – so kann z.B. eine große Topfpflanze, die auf einem Fußboden steht, sowohl als Hindernis in einer topologischen Karte der Umgebung, aber auch als manipulierbares Objekt (z.B. zum Gießen) gespeichert werden.

Natürlich muss der Roboter darüber hinaus noch weitere Informationen sammeln, die der Bewerkstelligung dieser Aufgabe dienlich sind – so benötigt der Roboter z.B.

[1] COGNIRON steht für „*cognitive robot companion*“. Siehe Internetseite des Projekts unter http://www.cogniron.org .

Informationen über seine Kommunikationspartner etc. Die Fähigkeiten, die der Roboter zum Sammeln der genannten Informationen einsetzen muss, bestehen insbesondere aus

- grundlegenden Dialogfähigkeiten
- Personenerkennung
- Gestenerkennung
- Raumwahrnehmung und Objekterkennung
- Navigation

Forschungen im Kontext des *home tour*-Szenarios fanden insbesondere an der Roboterplattform BIRON (*Bielefeld robot companion*) statt, die auch als technische Grundlage für diese Arbeit diente. Eine detaillierte Beschreibung des Aufbaus und der Fähigkeiten von BIRON wird in Kapitel 6.1 erfolgen.

Curious robot Im Gegensatz zu dem *home tour*-Szenario, in dem ein Roboter insbesondere durch menschliche Anleitung Informationen über seine Umgebung sammelt, ist der Roboter in dem *curious robot*-Szenario weitgehend auf sich allein gestellt. Der Roboter muss sich in einer Haushaltsumgebung orientieren und proaktiv über die Umgebung informieren. Da sich der Roboter dabei auch der Hilfe von anwesenden Personen bedienen darf, diese aber selbständig erkennen und ansprechen muss, könnte das Szenario als eine Erweiterung des *home tour*-Szenarios angesehen werden. Allerdings liegt im *curious robot*-Szenario der Schwerpunkt auf der eigenständigen Erkennung von Objekten und der Verwirklichung von „Neugier“ in einem Robotersystem.

Notwendige Fähigkeiten eines Roboters für das *curious robot*-Szenario sind:

- Objekterkennung
- Personenerkennung und -lokalisierung
- Fähigkeit zur Initiierung und Durchführung einer Interaktion mit Menschen
- Orientierung und Bewegung

Learning skills and tasks Das letzte Szenario des COGNIRON-Projektes beschäftigt sich mit dem Erwerb von Fähigkeiten und Aufgaben durch die Beobachtung von Menschen. Ein besonderer Schwerpunkt dabei ist die Erkennung von Zielen, denen Menschen während einer Handlung nachgehen.

Die notwendigen Fähigkeiten zur Bewältigung dieses Szenarios variieren mit den entsprechenden Aufgaben. Auf jeden Fall wird aber die Fähigkeit der Objekterkennung eine wichtige Grundlage darstellen.

Diese Szenarien decken einen Großteil der Anforderungen ab, denen ein Roboter-*companion* genügen müsste. Weitere Anforderungen würden zum Beispiel die Erweiterung und Entwicklung der sozialen Fähigkeiten des Roboters betreffen, z.B. die Einhaltung der korrekten Gesprächsdistanz zu einem menschlichen Kommunikationspartner (Walters u. a., 2005). Trotzdem sind mit diesen Szenarien Kernanforderungen formuliert, deren Erfüllung für einen Roboter-*companion* notwendig sind.

2.1 Natürliche Sprache als kommunikative Basis für Roboter-Companions

Diese Arbeit stellt – wie oben bereits beschrieben – den Versuch eines Beitrags zur Verbesserung der kommunikativen Fähigkeiten von Roboter-*companions* dar, wie sie von dem COGNIRON-Projekt entwickelt werden sollen. Wie leicht erkennbar ist, basieren zumindest die ersten beiden oben erwähnten Standardszenarien auf der Grundlage natürlichsprachlicher Kommunikation des Roboters mit Menschen. In der heutigen Alltagswelt jedoch ist solche Kommunikation von Menschen und Maschinen nur in wenigen Ausnahmefällen anzutreffen, je nach Definition sogar überhaupt nicht. Aus welchen Gründen benötigt ein Roboter-*companion* die Fähigkeit zur natürlichsprachlichen Interaktion, sogar erweitert um zusätzliche Modalitäten wie Gestik, wie aus dem Anforderungskatalog des *home tour*-Experiments ersichtlich ist?

Ich möchte versuchen in den folgenden Abschnitten eine Antwort auf diese Frage zu formulieren.

2.1.1 Komplexität als Problem angewandter Technologien

Mit der steigenden Komplexität von Geräten des alltäglichen Gebrauchs sinkt im Allgemeinen der Anteil der Benutzer, die die internen Funktionen des Gerätes verstehen und dasselbe nicht nur als *black box* verwenden. Ein gutes Beispiel dafür liefert die Automobilindustrie: Konnte noch vor wenigen Jahrzehnten ein einigermaßen geschickter Heimhandwerker mit dem entsprechenden Werkzeug sein Fahrzeug selbst reparieren, so werden heute moderne Automobile in bestimmten Fällen zu Spezialwerkstätten gebracht, da nur in diesen das entsprechende *know how* (und Werkzeug) existiert, mit dem Fehler repariert werden können. Selbst der unspezialisierte Automechaniker erfasst die einzelnen Automobile nicht mehr umfassend genug.

Auf der Ebene der Reparaturen von komplexen Geräten haben die Hersteller verschiedene Antworten gefunden, mit der steigenden Komplexität umzugehen: Bei mikroprozessorgesteuerten Geräten kann in den meisten Fällen einfach davon ausgegangen werden, dass die Software selbst fehlerhaft ist oder zur Kompensation von Fehlern des Geräts aktualisiert werden kann. Aus diesem Grund kann bei den meisten dieser Geräte auf relativ einfache Weise die geräteinterne Software (*firmware*) erneuert und so der Fehler mit geringem Aufwand behoben werden. Bei weniger kostspieligen Artikeln oder Teilen von Geräten wurde eine andere Lösung gefunden: Meist werden die Geräte oder Module selbst einfach ausgetauscht.

Zusätzlich zu dem Problem der Wartung und Reparatur komplexer Geräte tritt ein weiteres Problem auf, welches in den vergangenen Jahrzehnten starke Beachtung erhalten hat, die in der Zukunft nicht abnehmen wird: Das Problem der Benutzerfreundlichkeit oder *usability* eines technischen Geräts. Die ersten Computer waren Spezialgeräte, die nur von wenigen Personen verwendet und verstanden wurden. Die Art und Weise, wie ein Computer intern Probleme löste, stand in direktem Zusammenhang mit der Bedienung desselben; Programme wurden durch Schalter oder komfortabel durch Lochstreifen definiert und die Ausgaben erfolgten auf ähnlich umständliche Weise. Auch wenn es uns heute selbstverständlich scheint, dass technische Laien einen modernen Computer mit befriedigenden Resultaten bedienen können, so war es von den ersten Computern bis zu den modernen Desktop-Maschinen ein sehr langer

und mühevoller Prozess, bei dem prinzipiell immer eine Frage im Vordergrund stand: Wie kann ich die Interaktion mit dem Computer auf eine Art und Weise gestalten, dass sie dem Benutzer möglichst leicht fällt? Ein wichtiger Gesichtspunkt war die Abschirmung des Benutzers von den internen Funktionen des Gerätes durch *interfaces*, über die die Benutzer dem Computer auf eine angenehme Weise Befehle erteilen konnten.

Ideale *interfaces* waren in vielen Fällen solche, die für den Menschen „natürlich“ also an seine „Lebenswirklichkeit“ angelehnt waren. Ein *interface* stellt in dieser Hinsicht eine Kombination aus einem (oder mehreren) Ein- oder Ausgabegeräten und einem *code* dar, über den kommuniziert wird. Ein Beispiel für einen sehr weit von der Alltagserfahrung der Menschen entfernten *code* ist ein Lochstreifen, der ein Programm in Form von Steuerbefehlen kodiert. Auf der Ebene der Eingabegeräte war die Einführung der Tastatur ein großer Schritt in Richtung der Lebenswirklichkeit vieler Menschen, da sie den Umgang mit der alltäglichen Schriftsprache oder zumindest mit dem Alphabet ermöglicht. Auf der Ebene der *codes* stellten Hochsprachen wie C/C++ einen Schritt in die mathematisch geprägte Denkweise der Benutzer dar.

Heutzutage ist die Verwendung von grafischen Oberflächen in Form von *desktops* in Kombination mit Mäusen üblich. Auf der Ebene der Eingabegeräte entspricht die Maus einem Gegenstand, mit dem gezeigt bzw. gegriffen werden kann, auch wenn sie diese Wirkung nur in Verbindung mit dem *code*, einer virtuellen, aber an Schreibtische angelehnten grafischen Oberfläche erzielt (wofür noch ein visuelles Ausgabegerät benötigt wird). Auf letzterer ist es möglich, virtuelle Gegenstände zu bewegen und zu benutzen - nach meiner Ansicht ein weiterer Schritt in Richtung der Lebenswirklichkeit der Benutzer, sowohl auf der Ebene der *codes* als auch auf der Ebene der Eingabegeräte.

Als letztes Beispiel zeigt auch eine neuere Entwicklung, wieviel angenehmer der Umgang mit vertrauten Eingabegeräten und Modalitäten ist: Die im Jahr 2006 erschienene Spielekonsole „Wii“ verfügt über einen stabförmigen Controller, den man im dreidimensionalen Raum z.B. wie einen Tennisschläger, Golfschläger oder als Zeigegerät verwenden kann. Zumindest zum Zeitpunkt der Erstellung dieser Arbeit erfreute sich diese Konsole trotz der geringeren Rechenkraft im Vergleich zu anderen Konsolen sehr großer Beliebtheit.

2.1.2 Reduzierung der Bedienkomplexität durch natürlichsprachliche Kommunikation

Auch wenn die zur Zeit verbreitetste Art und Weise, mit einem Computer zu kommunizieren, aus einer Kombination aus Mausbewegungen und schriftlicher Kommunikation besteht, ist sie nicht die unter Menschen gebräuchlichste. Die „natürlichste“ Art und Weise, auf die Menschen miteinander kommunizieren ist Sprache, die von weiteren Modalitäten wie Gestik, Mimik, Prosodie etc. begleitet und unterstützt wird. Aus diesem Grund liegt es nahe, sowohl auf der Ebene der Eingabegeräte, als auch auf der Ebene des *codes* einen weiteren Schritt zu unternehmen. Das Resultat wäre die Verwendung von Lauten, Gesten, Mimik etc. (die durch ein Mikrophon etc. aufgenommen werden) zur Kommunikation in natürlicher, multimodaler Sprache (dem *code*).

Vor- und Nachteile natürlichsprachlicher Mensch-Maschine-Interaktion aus der Sicht des menschlichen Kommunikators und des Kommunikationsgegenstandsbereichs Wie auch bei den oben genannten Varianten quasi-natürlicher Kommunikation zwischen Mensch und Maschine würden sich aus der natürlichsprachlichen Mensch-Maschine-Kommunikation für die Seite des menschlichen Kommunikators gravierende Vorteile ergeben. Ich möchte an dieser Stelle darauf hinweisen, dass das Szenario einer typischen Mensch-Maschine-Kommunikation das einer mehr oder minder situierten Kommunikation ist, die als Ziel die Fertigstellung eines Arbeitsprozesses hat, also aufgabenorientiert ist. Ziele können das Erstellen eines Briefes, die Buchung einer Reise nach Skandinavien oder das Putzen eines Raumes sein - nahezu beliebige Aufgaben, für die quasi-intelligente Maschinen Anwendung finden können.

Mögliche Vorteile von natürlichsprachlich kommunizierenden Maschinen für Benutzer sind unter anderem:

(i) Verringerter bis gegen null gehender Lernaufwand für eine erfolgreiche und effektive Kommunikation

(ii) Verringerter Zeitaufwand für den Kommunikationsprozess

(iii) Gesteigerte Zufriedenheit durch quasi-soziale Interaktion

Sollte es sich bei dem Roboter um ein kommunizierendes System handeln, welches natürlichsprachliche Kommunikation tatsächlich umfassend beherrscht, so ist der erste Punkt wohl weitgehend unstrittig[2]: Die Benutzer müssen tatsächlich fast nichts lernen, um mit der Maschine zu kommunizieren, abgesehen einmal von Trivialitäten wie dem Namen des Gerätes.

Die Punkte (ii) und (iii) sind ggf. nicht unter allen Umständen zutreffend. Ich möchte das in den folgenden Absätzen verdeutlichen:

Ad (ii): Ob natürlichsprachliche Kommunikation ein optimales Medium zum Austausch über eine Aufgabenstellung ist, hängt nicht nur von den Fähigkeiten des Menschen oder der Maschine ab, sondern vor allem auch von der Aufgabenstellung selbst. Die Existenz von Formelsprachen in den Naturwissenschaften ist ein eindeutiges Zeichen dafür, dass bestimmte Sachverhalte nicht-natürlich weitaus effizienter kommuniziert werden können, als auf eine natürliche Weise.

Gerade aus dem Bereich der Informatik lassen sich viele Beispiele finden, in denen dies der Fall ist. So lässt sich eventuell ein Algorithmus zur Berechnung eines Wertes natürlichsprachlich nur schwer beschreiben, während eine Darstellung in einer Programmiersprache exakt und wahrscheinlich auch wesentlich kürzer als ein Dialog ist.

Fakt ist, dass ein Sprachsystem stets an eine Aufgabenstellung bzw. einen Raum, über den es zu kommunizieren gilt, angepasst ist. Diese Anpassung mag entweder durch noch nicht vollständig entschlüsselte Prozesse der Sprachentstehung/des Sprachwandels im Fall von natürlichen Sprachen – oder aber durch menschliches Design im Fall von Programmier- oder Formelsprachen – geschehen. Sogar natürliche Sprachen unterscheiden sich bereits stark nach der Lebensumgebung der Menschen, die sie entwickelt haben. Ein berühmtes Beispiel hierfür sind Eskimosprachen, die

[2] Ggf. kann durch die Notwendigkeit eines wechselseitigen *alignments* (Pickering und Garrod, 2004) ein gewisser Lernaufwand entstehen – das Argument ist folglich als pauschale Aussage zu verstehen, weniger als eine absolute.

eine Anzahl von Wörtern für „Schnee“ haben, die in den zweistelligen Bereich geht (Crystal, 1995, S.15). Diese Sprachen sind somit z.B. vom modernen Deutsch stark unterschieden, welches dafür etliche Pflanzensorten differenziert.

Die vorliegende Arbeit zielt wie beschrieben auf die Entwicklung von Kommunikationsfähigkeiten von intelligenten Maschinen, die sich im alltäglichen Lebensraum von Menschen befinden und über diesen mit ihnen kommunizieren müssen. Es ist sehr sicher anzunehmen, dass die Alltagssprache ein optimales Kommunikationsmedium zum Austausch von Informationen über diesen Raum darstellt, womit die gerade angebrachten Einwände nicht zutreffen. Ein Grund für diese Annahme beruht auf der wechselseitigen Anpassung von Sprache und Lebensraum, die alltäglich stattfindet.

Ad (iii): Forscher im Bereich der sozial interagierenden Roboter stellen immer wieder fest, dass sich sowohl Personen als auch Kulturen in Hinsicht auf den sozialen Umgang mit Robotern gravierend unterschieden. Ein künstliches Tier wie z.B. Paro (Wada u. a., 2002) kann sowohl als niedlich als auch als erschreckend angesehen werden. Personen behandeln Roboter freundlich oder geringschätzig, wie Menschen oder wie Maschinen. Durch die Existenz von Robotern entsteht anscheinend eine soziale Unsicherheit, die in Analogie zu anderen Kommunikationspartnern gefüllt wird – dies mag einer der Gründe sein, warum konkret an Haustiere angelehnte Roboter wie Sony's Roboterhund AIBO weniger soziales Unwohlsein auslösen als abstrakte Roboter. Realistische humanoide Roboter dagegen lösen das meiste Befremden aus: Der Grund dafür ist vermutlich, dass diese Roboter menschliche Kommunikationsformen durch ihr Aussehen nahelegen, dagegen aber in dem konkreten Prozess der Kommunikation die Fehlleistungen der Roboter um so stärker auffallen. Dieser Effekt ist als „**uncanny valley**“ u.a. in (Mori, 1982) beschrieben worden.

Das Ziel sollte also eine individuelle Anpassung der quasi-sozialen Interaktion an die jeweiligen kulturellen und persönlichen Umstände sein.

Einschränkungen aus der „Perspektive“ des künstlichen Kommunikators
Zusätzlich zu den genannten, aber eher geringfügigen Problemen existiert noch eine ganze Klasse weiterer, die tatsächlich in den vergangenen Jahren für die Forschung die mehr beachtete Herausforderung darstellten: Selbst wenn eine Kommunikationsform für den menschlichen Kommunikator angenehm erscheint und für den Gegenstandsbereich angemessen ist, kann sie jedoch den künstlichen Kommunikator vor ernste Schwierigkeiten stellen. So wurde in der Vergangenheit oft versucht, die Verarbeitung gesprochener Sprache (d.h., ohne Gestik, etc.) für den Computer fehlerarm zugänglich zu machen. Dabei stellten sich jedoch die folgenden Punkte als nicht trivial zu lösende Probleme heraus:

- Die Spracherkennung – also die Umsetzung von Schall in symbolisch codierte Worte – ist praktisch nie fehlerfrei handhabbar.
- Die korrekte Zuweisung von semantischen Analysen an Wortketten ist aufgrund der natürlichsprachlichen Ambiguitäten extrem schwierig.
- Das korrekte Verständnis von semantischen und pragmatischen Aspekten der Sprache ist aufgrund der damit einhergehenden Komplexität und Verankerung in die menschliche Gegenstands- und Wissenswelt für Computer nahezu unmöglich.

Viele der Probleme konnten in Einzelschritten näherungsweise gelöst werden, auch wenn sich stets herausstellte, dass ein natürlichsprachlich kommunizierendes System notwendigerweise die jeweiligen Verarbeitungsebenen von Sprache – als solche Ebenen werden Semantik, Pragmatik, Morphologie etc. angenommen – miteinander abstimmen muss, um maximal gute Ergebnisse zu erzielen. Da beispielsweise ein Spracherkennungssystem außerhalb von Laborbedingungen stets Erkennungsfehler produzieren wird, ist es auf ein Feedback z.B. von einer semantischen Analyse angewiesen, die hilft, die konkurrierenden Interpretationsmöglichkeiten des Spracherkenners neu zu bewerten. Auf diese Weise und höchstwahrscheinlich nur auf diese Weise können Dialogsysteme Ergebnisse erzielen, die nahe an menschlichen Kommunikationsfähigkeiten liegen.

Zusammenfassend kann die Frage nach dem Nutzen natürlichsprachlicher Systeme im Vergleich zu weniger „natürlichen“ Kommunikationsmodalitäten nur durch eine Funktion aus Aufwand und Nutzen beantwortet werden. Im Fall von situierter Kommunikation von Haushaltsrobotern mit Benutzern, die keinerlei technische Vorbildung besitzen, sollte jedoch als Daumenregel „je natürlicher desto besser“ angenommen werden, da das *uncanny valley* zumindest auf dem Gebiet der sprachlichen Kommunikation noch nicht erreicht ist und dem Benutzer – und ggf. auch dem Roboter – auf diese Weise weniger Lernaufwand entsteht.

„Natürlich“? – natürlich Bevor ich versuche, den Nutzen meiner Arbeit für natürlich kommunizierende Roboter-Dialogsysteme zu untersuchen, möchte ich an dieser Stelle parenthetisch den Begriff der „Natürlichkeit“ selbst genauer betrachten. Folgende Arbeitsdefinition entstand im Rahmen der Erstellung dieser Arbeit (Maas und Wrede, 2006):

Definition 2.1 (natürliche Kommunikation) *Eine Mensch-Roboter-Kommunikation ist* ***natürlich****, wenn*

1. *die Modalitäten der Kommunikation (Sprache, Gestik, Mimik etc.) dieselben sind wie für Vis-a-vis-Mensch-Mensch-Kommunikation, und*

2. *der menschliche Kommunikationspartner nicht lernen muss, mit dem Roboter zu kommunizieren.*

Der Bezug dieser Definition zum Nutzen natürlicher Mensch-Maschine-Kommunikation liegt auf der Hand; weiterhin gebe ich zu bedenken, dass diese Definition für (situierte) Mensch-Roboter-Kommunikation und nicht für Mensch-Maschine-Kommunikation im Allgemeinen gedacht ist.

Problematisch ist eventuell der zweite Punkt; man muss davon ausgehen, dass während jeder Kommunikation die Kommunikationpartner sich auf den jeweils anderen Kommunikator einstellen müssen, mindestens in Form eines permanenten *alignments* (Pickering und Garrod, 2004). Auf diese Weise findet ein stetiger Lernprozess statt, der auch in natürlichsprachlicher Kommunikation mit Robotern existieren würde. Tatsächlich bezieht sich der zweite Punkt auf die der Kommunikation vorgelagerten Lernprozesse wie das Lesen einer Anleitung, Durchführen eines Tutorials etc., die im Kontrast zu den in normaler Kommunikation integrierten Lernprozessen steht.

Eingeschränkte Kommunikation, also das Fehlen von wichtigen kommunikativen Fähigkeiten wie z.B. das Verständnis von Zeigegesten oder Ironie, würde zu längerer Einarbeitungs-/Lernzeit führen, da der Mensch erst die Einschränkungen des Robotersystems herausfinden und Vermeidungsstrategien entwickeln muss. Aus diesem Grund wurde diesem Aspekt kein weiterer Punkt gewidmet. Außerdem ist meiner Ansicht nach natürliche, situierte Kommunikation mit einem Robotersystem möglich, auch wenn dieses keine Ironie versteht - sofern die situative Gegebenheit nicht das Verständnis von Ironie erfordert.

Ein wichtiger Vorteil der Definition 2.1 ist, dass man aus ihr Grade an Natürlichkeit ableiten kann: Je mehr natürliche Modalitäten vorliegen oder je geringer die Einarbeitungszeit ist, als um so besser (im Sinne von natürlicher) kann die Kommunikation bezeichnet werden.

Nachdem ich den generellen Nutzen von natürlichsprachlichen *interfaces* im Bereich der situierten, aufgabenorientierten Mensch-Roboter-Kommunikation dargestellt habe, stellt sich natürlich die Frage, in wiefern die vorliegende Arbeit hilft, solche *interfaces* zu realisieren. Dieser Frage ist das folgende Kapitel gewidmet.

2.2 Themenerkennung als Unterstützung von natürlichsprachlichen Dialogsystemen

Eine wichtige Problematik, die speziell in den 80er Jahren in der KI[3] viel Beachtung fand, ist das mangelnde **Weltwissen** von Computern und Robotern, die mit der Umgebung kommunizieren sollen. So konnte z.B. ein Roboter mit Hilfe einer komplexen Software selbstständig berechnen, wie ein Turm aus Bauklötzen zu bauen sei – die Durchführung scheiterte jedoch zunächst an dem mangelnden Wissen des Roboters – das Gerät begann den Turm von oben zuerst aufzubauen, da es kein Wissen über Gravitation besaß.

Auch wenn in diesem Beispiel ein relativ einfacher Eingriff in das Programm das Problem behebt, steigt die Menge an benötigtem Hintergrundwissen bei komplexer Kommunikation oder Interaktion extrem an und ist nicht mehr durch einfache Tricks[4] zu implementieren. Da solche Mengen an Wissen üblicherweise nicht von Hand einprogrammiert werden können, sind in dieser Hinsicht selbstlernende Systeme zu bevorzugen, welche sich das fehlende Welt- und Kontextwissen aufbauend auf rudimentärem Lernalgorithmen selbst aneignen.

In dieser Arbeit möchte ich einen Ansatz vorschlagen, mit dem einem in einer situativen Kommunikation verankerten Dialogsystem die Möglichkeit gegeben wird, sich verhältnismäßig abstraktes dialogisches Wissen selbst anzueignen: Das Wissen über das aktuelle **Thema**, engl. *topic*. Dieses Wissen kann für einen natürlichsprachlich interagierenden Haushaltsroboter aus verschiedenen Gründen von Nutzen sein. Ich möchte im Folgenden kurz auf verschiedene mögliche Anwendungsfälle eingehen.

Ein wichtiges Anwendungsgebiet für thematische Information ist die Unterstützung des Sprachverstehens in Fällen, wo thematisches Wissen zum Verständnis vorausgesetzt wird: Bestimmte Fälle von **Anaphernresolution**, Referenzauflösung von Ob-

[3] KI: Abkürzung für **Künstliche Intelligenz**

[4] Im o.g. Fall die Regel: „Baue stets von unten nach oben!“

jekten im Allgemeinen, oder auch die Inferenz von Standardhandlungen, die innerhalb eines thematischen Kontexts mit Objekten verknüpft sind und auf die sprachlich Bezug genommen wird. Ein Beispiel wäre dafür die umgangssprachliche Formulierung „sich um das Geschirr kümmern“, was in den meisten Fällen vermutlich bedeutet, dass das Geschirr gereinigt wird. Andere mögliche Verwendungsgebiete für thematisches Wissen in einem natürlichsprachlich interagierenden Robotersystem wären zum Beispiel die Verbesserung der Spracherkennung durch thematische Einschränkung des Lexikons oder die verbesserte Strukturierung der Wissensdatenbanken des Roboters.

Im folgenden Beispiel kann eine sprachliche Referenz nicht ohne kontextuelles Wissen aufgelöst werden: Es handelt sich um einen fiktiven Dialog, in dem zwei verschiedene Themen, T1 und T2, existieren:

Beispiel 2.1
T1: Auf dem Tisch steht eine Pflanze. Die Pflanze ist grün, der Tisch braun. Der Tisch steht auf einem Teppich.

T2: Die Küchenzeile ist sehr klein. Neben der Spüle liegen eine Flasche mit Geschirreiniger, ein Schwamm und viele Teller. Außerdem ist dort noch eine Pflanze. Es ist eine Gewürzpflanze, Basilikum. [. . .]

Zu einem späteren Zeitpunkt wird der folgende Satz geäußert:
T?: Auf dem Teppich liegt ein Blatt von der Pflanze.

Ohne Kontextinformation lässt sich nicht folgern, welche Pflanze gemeint ist. Ein themenerkennendes System könnte jedoch – z.B. anhand der Erwähnung eines Teppichs – erkennen, dass es sich um Thema 1 handelt. Anschließend könnte durch eine einfache Inferenz geschlussfolgert werden, dass es sich um die Pflanze auf dem Tisch – und nicht die Gewürzpflanze – handelt. Selbstverständlich könnte dies auch über eine nicht auf dem Thema beruhende Inferenz gefolgert werden, so z.B. über den räumlichen Zusammenhang. Allerdings ist es wahrscheinlich, dass in ähnlichen Fällen themenbasierte Inferenzen weniger Fehler erzeugen würden; außerdem könnten Informationen verschiedener Systeme zu einem optimalen Ergebnis gekoppelt werden.

Ein anderes Beispiel wäre die oben erwähnte **Inferenz von Standardaktionen**: Wird in einem thematischen Kontext (Spüle) Geschirr stets mit der Handlung „Abwaschen“ in Verbindung gebracht, so kann diese als Standardhandlung in dem Satz „geh zu der Spüle und mach das Geschirr“ inferiert werden. Wenn hingegen der Satz „geh zu dem Kaffeetisch und kümmere Dich um das Geschirr“ geäußert wird, und Geschirr im Kontext des Kaffeetisches stets nur auf- und abgedeckt wird, kann diese Handlung zu einem anderen Thema gehörend ebenfalls inferiert werden.

Weitere mögliche Nutzen einer Themenerkennung für einen Roboter wären:

- Eine Verbesserung der **Spracherkennung** durch eine der Situation angepasste Modifikation des Sprachmodells (Lane u. a., 2003).
- Die Verknüpfung von **sozialem Verhalten** mit Themen. So kann der Roboter z.B. lernen, dass bestimmte Themen mit bestimmten Emotionen, Orten oder Zeiten einhergehen.
- Die **direkte Einbindung des Themas** zur Klärung der Gesprächssituation. So kann die Rückfrage des Roboters nach dem aktuellen Thema (bzw. die Äußerung desselben) sowohl Aufmerksamkeit simulieren, aber auch zum besseren Verständnis bei beiden Gesprächspartnern führen.

Bevor im nächsten Kapitel das Themenerkennungssystem schrittweise dargestellt wird, möchte ich im folgenden Abschnitt noch kurz auf den Gegenstandsbereich dieser Arbeit eingehen, nämlich die (nur scheinbar triviale) Fragestellung, was eigentlich ein Thema ist. Ich möchte dabei auch skizzieren welche Definitionen des Begriffs existieren, um so eine klarere Vorstellung davon zu geben, was mit dieser Arbeit im Speziellen erreicht werden soll.

2.3 Was ist eigentlich ein Thema?

„Yet the basis for the identification of 'topic' is rarely made explicit. In fact, 'topic' could be described as the most frequently used, unexplained term in the analysis of discourse.“
G. Brown und G. Yule[5]

Jeder informationstechnischen Arbeit, die es sich zur Aufgabe gemacht hat, ein bestimmtes Phänomen bzw. Muster in Daten zu erkennen oder zu analysieren, sollte nach Ansicht des Autors den Versuch einer möglichst genauen Definition des Gegenstandsbereichs unternehmen. In vielen Fällen ist dies verhältnismäßig einfach: So kann sich z.B. eine Arbeit zur Erkennung von Zellkernen in Photographien von Zellstrukturen auf die gängigen biologischen Definitionen von „Zelle“ und „Zellkern“ stützen. Eine Arbeit zur Erkennung von Themen in multimodaler Mensch-Maschine-Kommunikation sollte in Analogie hierzu in der Dialogforschung als einem Teilgebiet der Linguistik fündig werden können. Problematischerweise ist der Begriff *topic* – also Thema – von eher schwammiger Bedeutung, wie das diesen Abschnitt einleitende Zitat von Brown und Yule zeigt. Insbesondere da „Thema“ (*topic*) ein üblicherweise intuitiv und umgangssprachlich verwendeter Begriff ist, fällt eine exakte wissenschaftliche Definition, die der eigenen Intuition entspricht, schwer. Eine weitere Schwierigkeit bei der Findung einer einheitlichen Definition beruht auf der Fülle an Definitionen aus verschiedenen wissenschaftlichen Bereichen. Ich möchte diese verschiedenen Definitionen im Folgenden kurz thematisieren.

Satzthemen Ein vor allem in der Grammatiktheorie häufig verwendeter Themenbegriff ist das sogenannte **Satzthema** oder *sentential topic* (Chomsky, 1965) (Hoffmann, 1993). Wie der Begriff nahelegt, sind die Entitäten, die ein Thema innehaben, Sätze. Das Satzthema ist Teil eines Begriffspaares, welches in der neueren Forschung verschiedene Bezeichnungen erhalten hat: In der englischsprachigen Literatur findet sich meist die Unterscheidung zwischen *topic* und *comment*, in der deutschsprachigen Literatur wird meist das eingängige Begriffspaar **Thema** vs. **Rhema** verwendet.

Eine gängige Differenzierung der Begriffe geht z.B. auf Hockett (Hockett, 1958) zurück: Das *topic* eines Satzes wird von dem Äußernden zuerst angekündigt und liegt meist in Form des Subjektes vor, der *comment* ist dann die neue Information, meist in Form des Prädikats. Laut (Hoffmann, 2000) gehen die ursprünglich lateinischen Begriffe „Subjekt“ und „Prädikat“ in der Tat auf die altgriechischen Begriffe „Hypokeimenon“ („Vorliegendes“) und „Rhema“ („Gesagtes“) zurück. Die heute als grammatische Begriffe eingebürgerten Bezeichnungen werden also mit ihrer ursprünglicheren Bedeutung wieder in Bezug gesetzt.

[5](Brown und Yule, 1983, S.70)

Um ein Beispiel zu nennen:

Beispiel 2.2 *Maria ist gestern auf den Jahrmarkt gefahren.*

Dieser Satz enthält ein Subjekt, nämlich Maria. Das Subjekt wird zuerst genannt und dient der Ankündigung des Themas. Der Rest des Satzes enthält die neue Information und zählt nicht mehr als zum Thema, sondern zum *comment* oder Rhema gehörig.

Die Gleichsetzung von Subjekt mit Thema und Prädikat als *comment* ist allerdings in Ausnahmefällen nicht gegeben. Eine u.a. im Deutschen und Englischen häufig auftretende Ausnahme ist die so genannte **Topikalisierung**, bei der der *comment* in die Vorfeldstellung des Satzes und damit an die Subjektposition gehoben wird. Ein Beispiel aus (Hockett, 1958, S.201):

Beispiel 2.3 *That new book by Thomas Guernsey / I haven't read yet*

Das grammatische Thema des Satzes ist *„That new book by Thomas Guernsey"*, das Subjekt des Satzes ist jedoch das Personalpronomen „*I*". In der nicht-topikalisierten Version des Satzes, nämlich

Beispiel 2.4 *I haven't read that new book by Thomas Guernsey yet*

ist das Personalpronomen („*I*") sowohl das Subjekt, als auch das Thema des Satzes, was der Theorie der Gleichsetzung von Thema und *comment* widerspricht.

Wie schon beschrieben wurde in der Grammatiktheorie des öfteren versucht, eine Thema/Rhema-Definition zu finden, die sich direkt aus dem grammatischen Zusammenhang ergibt und eine Verbesserung der Thema/Subjekt und Rhema/Prädikat-Assoziation erlaubt. Ein prominenter Versuch stammt dabei von Chomsky, der in seinem revolutionären Buch „Aspects of the Theory of Syntax" (Chomsky, 1965) für die in diesem Buch dargestellte Grammatiktheorie eigens eine Konstituentenstellung im Satz entwickelte (ebd., S.221). Allerdings konnte dennoch in der Grammatiktheorie noch keine unbestrittene Definition des Begriffs Thema gefunden werden.

In der neueren Forschung ist die Diskussion von Thema und Rhema umfassend erweitert worden. Ein Problem der klassischen Definition entsteht z.B. durch die Annahme, dass das Thema eines Satzes Teil des Satzes ist. Das Problem wird u.a. am Beispiel der folgenden beiden Äußerungen eines (fiktiven) Dialogs deutlich:

Beispiel 2.5
(a) Hast Du gestern Maria gesehen?
(b) Nein.

„Maria" ist das Thema der zweiten Äußerung, die aber weder ein Satz ist, noch Maria explizit erwähnt. In neuerer Forschung finden sich daher grundsätzlich Definitionen von Thema und Rhema, die die Nennung des Themas nicht explizit fordern. Das Thema ist wie gehabt die bereits erwähnte, oder aber präsupponierte oder schon bekannte Information der Nachricht, das Rhema die neue. Das Thema kann somit implizit oder vorgenannt vorhanden sein. Sowohl Äußerungen/Sätze ohne Thema als auch Äußerungen/Sätze ohne Rhema sind möglich: Erstere treten oft am Anfang von Texten auf, letztere verstoßen gegen Konversationsmaximen, können aber somit pragmatisch zu interpretierenden Bedeutungsinhalt besitzen.

Die Thema/Rhema-Forschung ist noch nicht abgeschlossen, tatsächlich finden sich immer wieder neue Einwände gegen bekannte Definitionen. Ein Beispiel dafür findet man z.B. in der nachstehenden Satzfolge:

Beispiel 2.6 *Zahlreiche Zuschauer und Journalisten hatten sich eingefunden. Der Richter wies die Journalisten darauf hin, dass...*[6]

Das Substantiv „Journalisten“ aus dem 2. Satz ist ein Teil des Rhemas der eingebetteten Prädikation, obwohl die Journalisten vorerwähnt wurden. Ähnliche Probleme macht die genaue Definition von Thema und Rhema anhand von sprachlich eindeutigen Merkmalen.

Textthemen Im Kontrast zu der Definition von Thema als Thema eines Satzes, also eines *sentential topics*, kann der Begriff Thema auch über einen Text oder Diskurs definiert sein (z.B. das Thema von „Hamlet“). Solche Themen bezeichnet man als **Textthemen** oder **Diskursthemen** (*discourse topics*). Im Gegensatz zu Satzthemen ist der Träger eines Themas nicht ein Satz, sondern entweder ein Text oder Diskurs (Keenan und Schieffelin, 1976) oder das Thema findet seinen Träger in den Diskursteilnehmern selbst (Brown und Yule, 1983, S.68). Laut (Brown und Yule, 1983) geht die erste Unterscheidung von Satzthema und Textthema auf (Keenan und Schieffelin, 1976, S.380) zurück: Die Autoren wollten sich bewusst gegen die grammatischen Thema/Rhema-Definitionen abgrenzen und definierten den Begriff *discourse topic* nicht als eine Nominalphrase, sondern als eine Proposition (Aussage).

Die Annahme, dass für (fast) jeden Text eine zentrale Aussage existiert, die das Thema des Textes darstellt, ist letztendlich zu einfach gedacht, um zutreffend sein zu können – was wäre z.B. die zentrale Aussage von „Faust“? – auch wenn auf diese Weise ein Schritt zu einer intuitiveren Definition des Begriffs „Thema“ gegangen wurde. In vielen Fällen wird mittlerweile das Textthema als der inhaltliche Kern eines Textes bzw. als Antwort auf die Quaestio – also eine zentrale Frage der Art: „Was ist (dir) zum Zeitpunkt x am Ort y passiert?“ – des Textes gesehen (Klein und Stutterheim, 1992, S.3). Auch Formulierungen wie „*what is being talked/written about*“[7] sind nicht unüblich, stoßen aber auf das Problem, dass sich solche Themen nicht eindeutig formulieren lassen bzw. mehrere Formulierungen und damit sogar Sachverhalte existieren können. So könnte man das Thema eines Benutzerhandbuchs für eine Kaffeemaschine als die Kaffeemaschine selbst bezeichnen, oder als „(sicheres/einfaches/effektives) Arbeiten mit der Kaffeemaschine XY“ etc. Für weniger technische Textsorten nimmt das Problem rapide zu; das Thema von Shakespeares „Hamlet“ könnte man als „Drama um das tragische Ende eines jungen Adeligen“, „Leidensweg des Prinzen Hamlet“, „Intrigen um einen Königsmord in dem dänischen Königshaus“ etc. bezeichnen – all diese Antworten könnten aber zu Recht auf die Frage „Worum geht es in ‚Hamlet‘?“ gegeben werden.

Ich möchte an dieser Stelle betonen, dass die von mir vorgestellten Versuche, den Begriff „Thema“ zu definieren, eher einen exemplarischen Charakter haben. Sowohl das Gegensatzpaar Thema–Rhema, als auch der Begriff „Textthema“ werden weiterhin diskutiert. In der Tat scheint jeder Forschungsbereich, der sich mit dem Alltagsbegriff „Thema“ auseinandersetzt, eine eigene Arbeitsdefinition aufzustellen. Trotzdem hoffe ich, die spezifischen Unterschiede der Ansätze hervorgehoben zu haben.

[6](Bußmann, 2002, S.696)

[7](Brown und Yule, 1983, S.75), allerdings sind Brown und Yule keine Vertreter dieser Position

Ereignisbasierte Thema-Definitionen Eine drittes Forschungsgebiet, welches sich mit dem Begriff „Thema“ auseinandersetzt und aus welchem eine eigene Themendefinition hervorgegangen ist, ist der Bereich des *information retrieval*. Das Ziel dieser Disziplin ist eher locker definiert, ich möchte mich daher auf folgende Arbeitsdefinition festlegen: *Information retrieval*-Systeme (vgl. (van Rijsbergen, 1979) (Mehler, 2004)) analysieren eine Datengrundlage, um auf eine Anfrage (*query*) eine möglichst relevante Menge an **Dokumenten** aus einer semistrukturierten[8] Datenquelle zu liefern. Im Gegensatz zu Textmining-Ansätzen sind die Ergebnisse einer *information retrieval*-Anfrage höchstens nach der Relevanz sortiert, es kommt also nicht auf das Verhältnis der Dokumente zueinander an.

Eine wichtige Grundlage bzw. Spezialform des *information retrievals* ist natürlich die Sortierung nach Thema; angenommen, ein System verfügt über eine Datenbank von Zeitungsartikeln und soll auf die Frage, welche Artikel sich mit dem Mord an Präsident Kennedy beschäftigen, eine Antwort in Form von Dokumenten liefern. In diesem Fall kann ein automatisches Themenerkennungssystem die Antwort liefern.

Letztendlich eine Spezialform des *information retrieval* ist die Suche nach relevanten Dokumenten, die nicht in schriftlicher Form, sondern in Form von Audiodaten vorliegen. Audiodaten – wie z.B. Nachrichtenmitschnitte oder vorgelesene Texte – erlauben aufgrund der Problematik der Spracherkennung einen sowohl schwierigeren, als auch teilweise sehr unterschiedlichen Zugang zu der zugrundeliegenden Sprache. Ein Projekt, welches sich mit der thematischen Strukturierung von Audiodaten beschäftigt, ist das schon zuvor erwähnte Topic Detection and Tracking-Projekt. Im Kontext dieser Studie ist eine grundlegende Themendefinition entstanden ((Cieri u. a., 2002, S.42f), (Allan, 2002a, S.2), Übersetzung in das Deutsche von mir). Ich möchte sie im Folgenden exemplarisch darstellen.

Cieris Themenbegriff liegt der Begriff des Ereignisses (*event*) zugrunde:

Definition 2.2 (Ereignis (Cieri)) *Ein Ereignis ist eine bestimmte Entität, die zu einem spezifischen Zeitpunkt an einem spezifischen Ort geschieht. Zu dem Ereignis gehören weiterhin alle notwendigen Voraussetzungen und unmittelbaren (unavoidable) Konsequenzen der genannten Entität.*

Ein Beispiel für ein Ereignis ist zum Beispiel ein Attentat, welches zu einem bestimmten Zeitpunkt an einem bestimmten Punkt stattfand. Zu dem Attentat gehören auch die direkten Folgen, wie der Tod/die Verletzung des Opfers, das Entkommen/die Festnahme des Täters etc. Ich habe mir die Freiheit genommen, *unavoidable* (unvermeidlich) mit „unmittelbar“ zu übesetzten, da ich glaube, dass dies den Begrifflichkeiten im Deutschen näher kommt.

Die Definition des Themenbegriffs resultiert wie beschrieben aus der Definition des Ereignisbegriffs:

Definition 2.3 (Thema (Cieri)) *Ein Thema ist ein Ereignis oder eine Aktivität, sowie alle zu ihr/ihm in unmittelbarer Beziehung stehenden Ereignisse oder Aktivitäten.*

Ein nach Cieri zentraler Aspekt des von ihm vorgestellten Themenbegriffs ist die Zentriertheit auf konkrete Ereignisse, die im Gegensatz zum umgangssprachlichen Themenbegriff steht: In der Umgangssprache können allgemeine Ereignisse wie „Unfälle“

[8] Somit stehen *information retrieval*-Systeme im Gegensatz zu Datenbanksystemen (vgl. (Elmasri und Navathe, 2002)), die der effizienten Wiedergabe explizit gespeicherter Information dienen.

ebenso ein Thema sein wie ein konkreter Unfall, über den gesprochen wird. Diese Zentriertheit lässt sich meiner Ansicht nach wie schon bei den bisher dargestellten Themenbegriffen auf den Forschungskontext zurückführen, in dem die Definition stattfindet: Das TDT-Projekt ist als Teil der *information retrieval*-Forschung daran interessiert, aus natürlichsprachlichen Quellen Informationen zu gewinnen. Die dafür auszuwertenden Quellen müssen üblicherweise leicht verfügbar sein und Zugriff auf große Mengen von Daten liefern können, weswegen Nachrichten – sowohl Radio- und Fernsehnachrichten als auch Nachrichten in Printmedien – eine ideale Quelle darstellen. In Nachrichten wird meist über konkrete und selten über abstrakte Ereignisse berichtet, was die Besonderheit von Cieris Themenbegriff erklärt. Da die Definition von Cieri die Grundlage für eine thematische Annotation eines Nachrichtenkorpus darstellt, ist die Prägung des Begriffs leicht zu erkennen.

Unklar ist die Rolle des Wortes „Aktivität“ (*activity*); es kann sowohl in einem abstrakten Sinne (Marathonlaufen) als auch in einem konkreten Sinne (der Berliner Marathon 2005) definiert werden. Da Cieri nur Ereignisse als orts- und zeitgebunden definiert, aber über Aktivitäten nichts aussagt, interpretiere ich das Wort „Aktivität“ als eine Erläuterung des Ereignisbegriffs: Aktivitäten als Teilmenge der Ereignisse müssen also auch orts- und zeitgebunden sein, die Betonung von Ereignissen *und* Aktivitäten stellt nur eine Veranschaulichung dar.

Themendefinition dieser Arbeit Jeder der drei dargestellten Themenbegriffe lässt seine Herkunft klar erkennen; tatsächlich ist die Forschung von einer allgemeinen Definition des umgangssprachlichen Begriffs „Thema“ weit entfernt. Wie bei den meisten philosophischen Fragestellungen sollte dies jedoch kein Hinderungsgrund für die Entwicklung bzw. Verwendung einer spezifischen Themendefinition sein, man muss sich nur den Bezug der jeweiligen Definitionen zu den zugrundeliegenden Forschungsbereichen verdeutlichen.

Ich möchte nun auf die dieser Arbeit zugrundeliegende Definition des Begriffs „Thema“ zu sprechen kommen. Für diese Arbeit wird der Themenbegriff von Cieri – leicht angepasst – übernommen. Der Grund hierfür liegt klar in der gedachten Anwendung der zu klassifizierenden Themen: Satzthemen sind für die interne Steuerung eines Haushaltsroboters eher uninteressant bzw. werden durch das Sprachverstehen oder das Dialogsystem erkannt und dann (vermutlich) verworfen; von Interesse ist die Sammlung von Daten über allgemeinere Themenbereiche, als Satzthemen es sind. Folglich werden eher Diskursthemen gesucht; gegenüber den Definitionen von Diskursthemen bietet Cieris Definition jedoch den klaren Vorteil, direkt auf die Situiertheit der Kommunikation in Form von Ereignissen etc. eingehen zu können. Allerdings muss wie geschildert die Definition dazu geringfügig modifiziert werden. Dies geschieht in drei Schritten. Zuerst wird der Begriff des Ereignisses auf Ereignistypen bzw. wiederkehrende Ereignisse erweitert. Diese Erweiterung wäre gegebenenfalls auch für die Nachrichtendomäne des TDT-Projekts von Interesse, da nach der bisherigen Definition „Weihnachten 2006“ und „Weihnachten 2007“ keine thematische Einheit bilden – es sei denn, als „unmittelbare Konsequenz“ eines Ur-Weihnachts-Ereignisses[9], was aber letztendlich fragwürdig ist.

[9]Nach (Cieri u. a., 2002) existieren so genannte *seminal events* – also das Thema initiierende Ereignisse. Man erkennt den starken Bezug zu der Analyse von Nachrichten, bei der meist ein Ursprungsereignis zu einer Welle von ersten Nachrichtenbeiträgen führt, die sich dann fortsetzt

Definition 2.4 (Ereignis (diese Arbeit)) *Ein Ereignis ist eine bestimmte Entität, die zu einem spezifischen Zeit**intervall** an einem spezifischen Ort geschieht. Zu dem Ereignis gehören weiterhin alle notwendigen Voraussetzungen und unmittelbaren Konsequenzen der genannten Entität. Weiterhin stellen alle Ereignisse desselben Typs ein (abstraktes) Ereignis dar.*

Zu beachten ist weiterhin, dass der möglicherweise missverständliche Begriff „Zeitpunkt“[10] durch den naheliegenderen Begriff „Zeitintervall“ ausgetauscht wurde – Weihnachten 2005 fand an mehreren aufeinanderfolgenden Tagen und nicht an einem Zeitpunkt statt.

Durch die später folgende Erweiterung des Aktivitätsbegriffs ist diese Erweiterung des Ereignisbegriffs ggf. von eher untergeordneter Bedeutung – die Kommunikation mit einem Haushaltsroboter wird vermutlich eher über Aktivitäten als über wiederkehrende Ereignisse stattfinden. Auf diese Weise ist aber auch „Weihnachten“ oder „die Tagesschau“ ein mögliches Thema der Mensch-Roboter-Kommunikation.

In einem zweiten Schritt wurde für die situierte Kommunikation der Begriff **Objekt** hinzugefügt. Da Objekte Teil eines Ereignisses sind, konnten sie auch bei Cieri Teil eines Themas sein, allerdings nur in Verbindung mit einem Ereignis oder einer Aktivität. Bestimmte Gegenstände können aber in situierter Kommunikation selbst thematisiert werden, ohne dass dabei eine Handlung thematisch im Vordergrund steht. Ein Beispiel: Eine konkrete Blumenvase auf einem Tisch kann selbst thematisiert werden, indem über ihr Aussehen etc. gesprochen wird. Natürlich steht diese Vase stets in direkter Beziehung zu Aktivitäten (Riechen an den Blüten, Bewundern der Vase). Es scheint mir aber kontraintuitiv, diese als das eigentliche Thema anzusehen. So kann auf die Frage „worüber habt ihr gerade gesprochen?“ durchaus die Antwort: „wir haben uns über die Blumenvase auf dem Tisch unterhalten!“ gegeben werden, ohne dass der Fragesteller bei dieser Antwort die Stirn runzelt.

Definition 2.5 (Thema (diese Arbeit)) *Ein Thema ist ein (abstraktes) Ereignis, ein(e) Objekt(gruppe) oder eine Aktivität, sowie alle zu ihr/ihm in unmittelbarer Beziehung stehenden Ereignisse, Objekte oder Aktivitäten.*

Ein weiterer gravierender Unterschied zu Cieris Definition ist die Definition von „Aktivität“. Im Gegensatz zu der von mir bei Cieri angenommenen Definition möchte ich Aktivitäten doch als Abstraktum definieren. Cieris (unterstellten) Aktivitätsbegriff möchte ich in dem Begriff „Ereignis“, den ich von Cieri in großen Teilen übernehme, subsumieren.

Definition 2.6 (Aktivität (diese Arbeit)) *Eine Aktivität ist eine einmalige oder wiederkehrende Handlung, samt aller in direktem Zusammenhang stehenden Objekte, Aktivitäten und/oder Ereignisse.*

Ein Beispiel für eine Aktivität in situierter Kommunikation könnte z.B. „Staubsaugen“ sein. Zum Oberbegriff des Staubsaugens gehören u.a. alle Ereignisse der Handlung „Staubsaugen“, Staub und natürlich der Sauger.

Die Vorteile dieser Definition gegenüber der Definition 2.2 liegen in der weitaus größeren Flexibilität – viele Themen aufgabenorientierter, situierter Kommunikation

oder beendet.

[10] Herzlichen Dank an Alexander Mehler für den Hinweis.

können wie oben beschrieben Objekte, aber auch abstrakte Ereignisse und Handlungen als Thema besitzen. Der größte Nachteil der Definition liegt in dem intuitiv zu interpretierenden Begriff der unmittelbaren Beziehung. Cieri umgeht dieses Problem, indem für jede Gruppe von Ereignissen eines konkreten Korpus (d.h., Ereignisse in unmittelbarer Konsequenz eines *seminal event*) eine Gruppe von Interpretationsregeln definiert wird, anhand derer die Annotatoren mehr oder weniger klar entscheiden können, welche anderen Ereignisse zu dem Thema zu zählen sind. Diesen Luxus kann sich eine allgemeinere Definition natürlich nicht leisten; außerdem ist es naheliegend, dass die Entstehung von Themen durch die jeweilige Situation bzw. Umwelt geprägt ist, wodurch das Konzept der unmittelbaren Beziehung situations- bzw. aufgabenabhängig wäre. Da aus diesem Grund eine genauere Definition sehr wahrscheinlich nicht möglich bzw. nicht aussagekräftig ist, möchte ich an dieser Stelle auf die Intuitionen der Leser verweisen.

Nachdem ich für diese Arbeit eine Arbeitsdefinition des Begriffs „Thema“ vorgeschlagen habe, möchte ich in dem folgenden Kapitel die theoretischen Vorüberlegungen, die zur Struktur des Themenerkennungssystems geführt haben, darstellen.

3 Modellierung des Offline-Themenerkennungssystems

In diesem Kapitel sollen die Grundlagen so wie der grundsätzliche Aufbau des erstellten Themenerkennungssystems skizziert werden. Die genauen Details der Implementierung und der Integration in das Robotersystem werden dagegen in den jeweiligen Kapiteln 5 und 6 dargestellt.

Eine Themenerkennung auf einem dialogfähigen Robotersystem kann üblicherweise auf viele verschiedene roboterinterne Informationsquellen zurückgreifen, die sich durch die Vorverarbeitung von Sensordaten erschließen lassen. In manchen Fällen kann davon ausgegangen werden, dass die Vorverarbeitungsschritte in irgendeiner Form schon auf dem Robotersystem realisiert sind. Ein Beispiel dafür wäre die Spracherkennung, also die Umwandlung von Audiosignalen in schriftlich kodierte Sprache, ohne die ein Robotersystem nicht in der Lage wäre, natürlichsprachlich zu kommunizieren. Allerdings ist selbst diese Annahme potentiell kritisch und gilt im Wesentlichen nur für Robotersysteme, die für möglichst komfortable – und daher „natürliche“ Mensch-Maschine-Kommunikation konzipiert wurden.

Limitierend auf die Art und Anzahl der Informationsquellen wirkt sich natürlich die Ausstattung des Roboters mit Sensoren aus. Zwar verfügen die meisten Roboter über eine Kamera und Mikrophone, aber andere Sensoren wie Laser-Range-Scanner, Ultraschallsensoren oder Wärmebildkameras stellen keine Standardausrüstung dar. Auch die physikalischen Unterschiede zwischen den Sensoren, wie z.B. Empfindlichkeit oder Stereofoniefähigkeiten bei Mikrophonen können unter Umständen gravierend sein, so dass auch die Qualität der Informationsquellen stark variabel ist. Zuletzt wirkt sich natürlich die Speicher- und Rechenkapazität der Roboter-Computer möglicherweise begrenzend auf dessen Fähigkeiten zur Echtzeit-Vorberechnung aus.

Die Frage nach den Voraussetzungen für eine Themenerkennung auf einem mobilen Roboter kann aufgrund der geschilderten hohen Variabilität letztendlich nur in Bezug auf eine konkrete Roboterplattform in Kombination mit der dazugehörigen Softwareumgebung gestellt werden. Diese Aussage mag etwas pessimistisch klingen, jedoch ist trotzdem davon auszugehen, dass die Entwicklung eines Themenerkennungssystems für einen bestimmten Roboter eine solide Grundlage für die Entwicklung eines solchen Systems zumindest auf verwandten Robotersystemen darstellt. Nur die konkrete Implementation und/oder die Anzahl von Vorarbeiten, die von dem Themenerkennungssystem oder dem Roboter übernommen werden müssen, werden starke Unterschiede aufweisen, jedoch nicht die Struktur des Ansatzes selbst.

Im Rahmen dieser Arbeit wurde ein Themenerkennungssystem für den Roboter BIRON geplant und erstellt. Die Wahl von BIRON ergab sich u.a. aus der Wahl des Testszenarios (*home tour*), für welches die Software von BIRON speziell konzipiert wurde. Auf die Gründe, dieses Szenario zu wählen, wird in Kapitel 4.1 auf Seite 73 genauer eingegangen werden. Weitere Gründe für die Wahl von BIRON bestanden

natürlich in der Verfügbarkeit des Systems und dem verhältnismäßig sehr weiten Entwicklungsstand, der viele für eine Themenerkennung nützliche Daten und Vorverarbeitungsprozesse zu liefern in der Lage war. Die Hard- und Softwareausstattung von BIRON wird im Kontext der Diskussion der Implementierung im Abschnitt 6.1 auf Seite 123 dargestellt werden.

Die (teilweise noch zu realisierenden) Fähigkeiten von BIRON definierten den Rahmen in Form von verfügbaren Informationsquellen, auf die das entwickelte Themenerkennungssystem Zugriff haben kann. Im Wesentlichen sind dies ein Spracherkennungssystem, ein Dialogsystem, ein System zur Analyse der Aufmerksamkeitszustände des Kommunikationspartners so wie eine (Wieder-)Erkennung von in die Kommunikation einbezogenen Objekten. Das entwickelte Verfahren wurde aber so allgemein wie möglich gehalten, so dass auch eine vergleichbare Implementierung auf einem alternativen Robotersystem potentiell möglich ist.

In dem vorliegenden Kapitel soll weniger eine Darstellung des Themenerkennungssystems aus der „Vogelperspektive“ versucht werden, als vielmehr eine schrittweise Nachvollziehung der Gründe und Überlegungen, anhand derer sich das System zu seinem momentanen Stand entwickelt hat. Aus diesem Grund beginnt das Kapitel mit einem Überblick über relevante Vorarbeiten.

3.1 Vorarbeiten

Die Erkennung von Themen in situierter HRI (*human robot interaction*) ist ein neues Feld, welches bisher noch nicht untersucht wurde. Auf dem Gebiet der Themenerkennung wird jedoch schon seit vielen Jahren geforscht. Im Folgenden werden insbesondere die geistigen Vorläufer dieser Arbeit kurz beschrieben. Die nachfolgenden Abschnitte in diesem Kapitel dienen dann der genauen Diskussion spezifischer Herangehensweisen, die für diese Arbeit Verwendung fanden.

3.1.1 Information retrieval

Wie bereits weiter oben erwähnt (vgl. Abschnitt 2.3 auf Seite 21) fanden viele Forschungen zur Themenerkennung im Bereich des *information retrieval* (vgl. u.a. (van Rijsbergen, 1979)) statt. Die vorliegende Arbeit ist stark von diesen Arbeiten beeinflusst worden, wobei im Allgemeinen aber eher grundsätzliche Herangehensweisen (Lemmatisierung, Verwendung statistischer Verfahren) eingeflossen sind. Im Speziellen wurde diese Arbeit insbesondere beeinflusst durch die Forschung im Bereich der so genannten semantischen Räume. Diese wurden bisher u.a. dazu verwendet, um die automatische Verschlagwortung von Dokumenten (*indexing*) zu unterstützen. Auf die jeweiligen Modelle (einfaches Vektorraummodell, Fuzzy Semantics und Latent Semantic Analysis) wird an späterer Stelle (Abschnitt 3.4) detailliert eingegangen.

3.1.2 Textsegmentierung

Eng mit der Aufgabe des *information retrieval* (und damit mit der Forschung im Gebiet der Themenerkennung) verknüpft ist das Forschungsgebiet der **Textsegmentierung**. Dies scheint auf den ersten Blick nicht plausibel zu sein, wird aber verständlich, wenn man realisiert, dass lineare Textsegmentierung darauf zielt, ein Dokument

(einen Text) in Blöcke zu unterteilen, die in sich kohärent sind und bei denen aufeinanderfolgende Blöcke unterschiedliche Themen besitzen (Choi u. a., 2001). Der Hauptunterschied zwischen einer dynamischen Themenerkennung in einem Fließtext und einer Textsegmentierung scheint darin zu bestehen, dass die Themenerkennung zusätzlich versucht, Informationen über das jeweilige Thema zu gewinnen – also z.B. auch zu erkennen, wann ein Thema wiederkehrt –, während die Aufgabe der Textsegmentierung mit einer reinen Unterteilung abgeschlossen ist. Dies hat aber starke Auswirkungen auf die zu verwendenden Algorithmen. Um ein Beispiel zu nennen: In dem bekannten *TextTiling*-System von Marti Hearst (Hearst, 1997) kommen zwei Basisalgorithmen zum Einsatz: Blockvergleich (Hearst und Plaunt, 1993) und die Analyse der Einführung neuen Vokabulars (Youmans, 1991). In beiden Verfahren wird ein Fenster von benachbarten Sätzen (Blockvergleich) bzw. Worten (Vokabeleinführung), die für den Algorithmus im aktuellen Schritt sichtbar sind, auf den Text gelegt, welches schrittweise verschoben wird. Innerhalb dieses Fensters wird dann die Anzahl an übereinstimmenden Begriffen bzw. die Anzahl neuer Wörter gemessen und in Funktionswerten ausgedrückt. Die Kohäsion des Textes lässt sich anhand des entstehenden Funktionsverlaufs über dem Text ablesen. Wie jedoch ersichtlich ist, treffen beide Verfahren keine Aussage über die Art der Segmente. Diese ließe sich höchstens über parallel gesammelte Informationen – z.B. anhand einer *history* der neu eingeführten Wörter – und einen anschließenden Vergleichsprozess der ermittelten Segmente ermitteln.

Trotzdem entspricht die grundsätzliche Struktur des in dieser Arbeit entwickelten Systems in vielen Aspekten dem in (Choi u. a., 2001) skizzierten Basisaufbau eines Themensegmentierungssystems. Dabei werden drei Hauptaufgaben solcher Systeme differenziert:

1. Erkennung elementarer Blöcke: Der lineare Text wird in thematisch elementare Blöcke unterteilt, die keinen Themenwechsel beinhalten.

2. Ähnlichkeitsmaß: Die Ähnlichkeit zwischen den gebildeten Blöcken wird anhand eines Ähnlichkeitsmaßes bestimmt.

3. Clustern: Anhand dieser Ähnlichkeit werden die Blöcke zu Themenabschnitten verschmolzen.

Diese Aufgaben werden auch von dem in dieser Arbeit entwickelten System erfüllt, eine detailliertere Herleitung derselben in Bezug auf diese Arbeit findet sich in den diesem Unterkapitel folgenden Abschnitten.

In (Choi u. a., 2001) findet eine Unterscheidung in zwei Typen von Verfahren zur Textsegmentierung, nämlich textkohäsionsbasierte und solche, die auf verschiedenen Quellen (Modalitäten) basieren, statt:

Textkohäsionsbasierte Verfahren Diese Verfahren gehen historisch auf (Halliday und Hasan, 1976) zurück. In dieser Arbeit wurde die Kohäsion über wiederkehrendes Vokabular bestimmt; Textregionen mit geringer Textkohäsion – also Textabschnitte, in denen das Vokabular wechselt – entsprechen Themenwechseln. Nachfolgende Arbeiten modifizierten oftmals die Art der einem Textabschnitt zugrundeliegenden Entitäten, deren Vorhandensein Kohäsion bzw. die Abwesenheit derselben kennzeichnet. So ist ein häufig begangener Weg die Reduktion der grammatischen Information

von Worten durch Stammformenbildung bzw. Lemmatisierung, oder aber die Abbildung von Worten/Phrasen auf semantische Entitäten. Dies kann z.B. durch Thesauren geschehen, wie in (Morris und Hirst, 1991) bzw. (Morris, 1988) oder aber durch die Projektion der Blöcke in (latent) semantische Räume (Choi u. a., 2001) (Bestgen, 2006). Die Idee hinter diesen Verfeinerungen ist, dass Textkohäsion primär ein semantisches Phänomen ist, welches auf der syntaktischen Ebene nur unzureichend reflektiert wird. Die Projektion eines Textes auf semantische Entitäten bzw. Relationen – selbst wenn diese mit Hilfe der syntaktischen Ebene gewonnen wird – kann so hilfreich für das Erfassen von Kohäsion sein.

Im Rahmen dieser Arbeit wird das Verfahren des Erfassens von Textkohäsion durch Projektion von Texten (Diskursen) in semantische Räume übernommen. Dieses Verfahren hat im Vergleich zu den beschriebenen Verfahren des Blockvergleichs bzw. der Analyse neuen Vokabulars den Vorteil, dass gleichzeitig verhältnismäßig einfach Informationen über die jeweiligen Themen gesammelt werden können. Detailliertere Informationen finden sich dazu u.a. in Abschnitt 3.4.

Multi-Quellen-Verfahren Im Gegensatz zu den rein textbasierten Verfahren existieren noch solche, die zusätzliche Informationen als gegeben betrachten und verwenden können. Meist handelt es sich dabei um Daten aus Audioquellen, aber auch Videoquellen sind nicht unüblich. Typische Informationsquellen neben den schon genannten sind die Erkennung von Schlüsselphrasen, Prosodische Merkmale (Shriberg u. a., 2000), Pausenmodelle, energiebasierte Ansätze (Greiff u. a., 2000) etc. Üblicherweise wird eine große Anzahl von Merkmalen generiert und dann anhand eines Trainingskorpus eine Gewichtung derselben für die endgültige Entscheidungsfindung vorgenommen. So existieren z.B. entscheidungsbaumbasierte Verfahren (Litman und Passonneau, 1995) (Shriberg u. a., 2000), probabilistische Modelle (Hajime u. a., 1998) und Maximum-Entropie-Modelle (Beeferman u. a., 1997) zur Gewichtung der Parameter.

Meiner Ansicht nach ist es problematisch, diese Verfahren als Verfahren zur Textsegmentierung zu klassifizieren, da sie in vielen Fällen tatsächlich zur Videosegmentierung, Audiosegmentierung o.ä. herangezogen werden und die genannten Merkmale für den größten Teil aller Texte nicht zur Verfügung stehen. Ihr grundsätzlicher Nutzen für diese Arbeit steht natürlich nicht in Frage, da in situierter HRI nahezu immer mehrere Quellen (Audio, Video) verfügbar sind.

Die fundamentale Datenquelle für das im Rahmen dieser Arbeit zu entwickelnde System ist – wie oben schon angedeutet – natürliche Sprache, vorliegend in Audiostreams. Die wahrscheinlich umfassendste Forschung zum Erkennung von Themen in Audiodaten, die gesprochene Sprache enthalten, fand im Rahmen des US-amerikanischen Topic Detection and Tracking-Projektes (TDT) statt. Ich möchte in den folgenden Abschnitten detailliert auf das TDT-Projekt und den im Rahmen dieses Projektes entwickelten Ansatz zur Themenerkennung eingehen.

3.1.3 Das TDT-Projekt

Das TDT (Topic Detection and Tracking)-Projekt ist ein DARPA-finanziertes Forschungsprojekt verschiedener amerikanischer Universitäten. Das Projekt hat es sich zur Aufgabe gemacht zu erforschen, wie aus unsegmentierten Sprachdatenströmen

verschiedener Sprachen ohne menschliche Hilfe – also nur durch maschinelle Verarbeitung – die thematischen Strukturen extrahiert werden können. Ein Anwendungsszenario ist eine thematische Suche auf großen Mengen von Radionachrichten, innerhalb derer für einen an einem Thema interessierten Hörer auf diese Weise eine Vorauswahl getroffen werden soll.

Datengrundlage

Das Datenmaterial des TDT-Projektes liegt in verschiedenen, mehrsprachigen Korpora vor. Allen ist gemeinsam, dass sie ausschließlich auf authentischem Nachrichtenmaterial beruhen. Grundsätzlich wurden alle Korpora entweder in Textform aufgezeichnet oder manuell oder automatisch transkribiert. Die schon in Textform vorliegenden Teile der Korpora stammten direkt von Nachrichtenagenturen; so besteht ungefähr die Hälfte des ca. 16.000 Artikel umfassenden Pilot-Korpus des TDT-Projektes aus in Textform vorliegenden Nachrichten der Agentur Reuters. Die zweite Hälfte des Korpus besteht aus Nachrichten des Senders CNN, die manuell transkribiert wurden. Dass ein bedeutender Teil des Korpus nicht in einem Audioformat vorliegt, erklärt sich durch die großen Anforderungen, die eine manuelle Transkription mit sich bringt: Auf diese Weise konnte den Teilen des Projektes, die nicht notwendigerweise mit den Audiodaten arbeiten müssen, eine größere Datenmenge zur Verfügung gestellt werden. Trotzdem besteht die Zielsetzung des Projekts in der Themenerkennung auf Audio- und nicht auf Textdaten.

Grundsätzlich wurden alle TDT-Korpora manuell nach vorgegebenen Themen annotiert – am Beispiel des Pilot-Korpus waren dies 25. Die Annotation unterlag festgelegten Richtlinien, ab wann ein Nachrichtenbeitrag als zu einem Thema gehörig zählte.

Basisaufgaben

Im Rahmen des TDT-Projektes wurden Themenerkennungsprozesse im Allgemeinen in eine Anzahl von fünf Basisaufgaben unterteilt. Jedes Basisszenario – mit Ausnahme der *story segmentation* – steht für ein Anwendungsgebiet von Themenerkennungssystemen. *Story segmentation* ist dagegen eine Voraussetzung für die anderen vier Basisszenarien.

Die Basisaufgaben sind im folgenden:

1. *story segmentation*
2. *first story detection*
3. *cluster detection*
4. *tracking*
5. *story link detection*

Ich möchte die folgenden Abschnitte dazu verwenden, genauer auf die genannten Kernaufgaben einzugehen. Meine Darstellung folgt dabei im Wesentlichen der zusammenfassenden Darstellung aus (Allan, 2002a).

Story segmentation Die Aufgabe der *story segmentation* (Segmentierung) ist es, einen Nachrichten- bzw. Datenstrom in einzelne Segmente zu unterteilen, die jeweils ein Thema besitzen. Somit ist sie mit der in Abschnitt 3.1.2 auf Seite 27 diskutierten Aufgabe der Erkennung elementarer Blöcke identisch.

Üblicherweise orientiert man sich bei der Analyse von Radio- oder Zeitungsnachrichten an der von der Nachrichtenquelle vorgegebenen Struktur: So stellt z.B. ein Zeitungsartikel eine *story* dar, die dann als Trägerin eines Themas angesehen wird.

Es ist unmittelbar einsichtig, dass die Lösung der Aufgabe der Segmentierung stark abhängig von den erhältlichen Daten ist. Das TDT-Projekt beschäftigt sich insbesondere mit Nachrichtendaten aus Zeitungen oder Radioquellen. Die Gliederung von Zeitungsartikeln ist in den meisten Fällen schon im Datenmaterial verankert, die Segmentierung ist daher schon gegeben oder – wenn die Artikel in elektronischer Form gespeichert vorliegen – trivial. Forschung zur Segmentierung von Nachrichtendaten aus Zeitungen wurde im Rahmen des TDT-Projektes somit nicht durchgeführt. Laut (Allan, 2002a) basierten im TDT-Projekt die Versuche, Audionachrichten zu segmentieren, primär auf dem transkribierten Korpus und weniger auf den reinen Audiodaten, obwohl auch solche Ansätze in der Vergangenheit vielversprechende Ergebnisse geliefert haben (Shriberg u. a., 2000) (Stolcke u. a., 1999).

Die Ansätze, die sich mit den transkribierten Daten auseinandersetzten, unterteilen sich in vokabel- bzw. konzeptbasierte Ansätze (van Mulbregt u. a., 1999) (Ponte und Croft, 1997) und Ansätze, welche die Grenzen von Nachrichtenbeiträgen anhand von Schlüsselwörtern, Pausen o.ä. feststellen (Beeferman u. a., 1999). Wie geschildert wird an späterer Stelle auf diese Basisaufgabe eingegangen.

First story detection *First story detection* oder FSD beschäftigt sich mit der Erkennung der ersten Nachricht, die zu einem neuen Thema gehört. Stellt man sich die Themenerkennung auf einem Nachrichtendatenstrom vor wie eine Person, die auf einem Fließband ankommende Objekte nach Typ in dafür vorgesehene Behälter einsortiert, so ist es notwendig, diese Person über neue Objekttypen zu informieren – und einen neuen Behälter aufzustellen. FSD ist somit – wie die Segmentierung – eher eine Voraussetzung von anderen Themenerkennungsszenarien, insbesondere der *cluster detection*. Allerdings kann sie auch für sich genommen Nutzen bringen, so z.B. in einem System, welches speziell über neue, unbekannte Ereignisse informieren soll.

Ansätze zur FSD vergleichen üblicherweise neu eingehende Nachrichten mit den schon verarbeiteten Nachrichten. Dazu werden Merkmale extrahiert, die üblicherweise auf Wortverteilungen basieren (Allan u. a., 1999) (Allan u. a., 2000) (Jin u. a., 1999). Unterschreitet die Ähnlichkeit einen bestimmten Schwellwert, wird angenommen, dass die neue Nachricht zu einem unbekannten Thema gehört.

Cluster detection Die Aufgabe der *cluster detection* impliziert die Aufgabe der FSD. Ein *cluster detection*-Algorithmus gleicht einer Person, die über ein Fließband Objekte ihr unbekannten Typs geliefert bekommt. Die Person muss diese Objekte nach eigenem Ermessen in unbeschriftete Kisten einsortieren, wobei stets (thematisch) ähnliche Objekte in einer Kiste zusammenkommen sollen. Die Grundannahme ist somit, dass *cluster detection*-Systeme unüberwacht lernen sollen, wobei ihnen nicht einmal die Anzahl möglicher Themen bekannt ist. Die sortierende Person muss sich also selbst um neue Container bemühen.

Die Aufgabe der *cluster detection* ist die wesentliche Basisaufgabe, die im Rahmen dieser Arbeit für ein natürlichsprachlich kommunizierendes Robotersystem gelöst werden musste. Ich werde zu einem späteren Zeitpunkt noch einmal genauer auf diese Aufgabenstellung eingehen.

Tracking Diese Aufgabe ist eng verwandt mit Fragestellungen aus dem klassischen *information retrieval*. Zu einer sehr kleinen Menge von gegebenen Nachrichten sollen alle neu hinzukommenden Nachrichten desselben Themas gefunden werden. Ein Anwendungsbeispiel hierfür wäre ein Autoradio, welches z.B. nur die Sportnachrichten abhört, oder aber ein Informationssystem, welches aus den im Internet verfügbaren News-Tickern für einen Börsenmakler relevante Wirtschaftsnachrichten findet. Aufgrund der Nähe zum *information retrieval* konnten Forschungen im Bereich Tracking in der Vergangenheit relativ schnell mit sehr guten Ergebnissen aufwarten. Übliche Verfahren berechnen die Ähnlichkeit von (meist vektoriellen) Repräsentationen zu klassifizierender Dokumente zu bekannten, thematisch vorklassifizierten und ermitteln auf diese Weise die thematische Zugehörigkeit.

Story link detection Hinter diesem Begriff steht die Aufgabe, durch aus Daten erhaltenes Wissen erkennen zu können, ob zwei Nachrichten dasselbe Thema haben. Laut (Allan, 2002a) ist der Nutzen einer Lösung dieser Aufgabe nicht unmittelbar einsichtig; tatsächlich stellt es wiederum eine Grundlage für die anderen Aufgabengebiete dar, obwohl ein perfektes System zur *story link detection* Tracking, FSD und *cluster detection* (und natürlich auch Segmentierung) beherrschen müsste (vgl. (Allan, 2002a, S.7 Z.19ff)). Diese Aussage erscheint mir etwas unbefriedigend, da meiner Ansicht nach die Aufgabe der *story link detection* (SLD) vielmehr nur eine Umformulierung und Verschärfung der Tracking-Aufgabe ist. Auch erscheint das Argument zirkulär: Wenn SLD eine Kerntechnologie für die anderen Aufgaben ist, warum soll diese Kerntechnologie dann selbst auf die genannten Technologien zurückgreifen müssen?

Eine Umformulierung der Aufgabe der SLD scheint sinnvoll: So könnte die Forderung nach einem datengetriebenen Ansatz dahingehend reduziert werden, dass SLD-Algorithmen nur einen rein oberflächlichen Vergleich von *story*-Paaren durchführen sollen. Im Rahmen dieser Arbeit ist jedoch die Aufgabe der SLD eher unwichtig, so dass im Folgenden diese theoretischen Überlegungen zurückgestellt werden.

Anhand der Differenzierung der Basisszenarien möchte ich im Folgenden die Struktur des entwickelten Systems genauer betrachten.

3.2 Struktureller Ansatz

Im Einleitungskapitel wurden verschiedene Anforderungen an ein Themenerkennungssystem auf einem mobilen Roboter geschildert. Diese waren im Speziellen:

- die Fähigkeit, Themen dynamisch zu bilden, um sich an schnell wechselnde Umgebungen anpassen zu können (dynamisch)
- die Fähigkeit, das aktuelle Thema möglichst ohne Zeitverzögerung zu erkennen (*online*)

- die Fähigkeit, zur Verbesserung der Themenerkennung die auf einem Robotersystem vorliegende zusätzliche Information verwenden zu können (Multimodalität)
- die Fähigkeit, auch in aufgabenorientierter Kommunikation Themen zu erkennen

Auf die letzte Anforderung wird im Rahmen dieser Arbeit nicht mehr eingegangen; die im Rahmen dieser Arbeit zu lösende Aufgabe ist die Bewerkstelligung des *home tour*-Szenarios.

Ein weiterer Punkt erscheint trivial, soll aber dennoch betont werden:

- Der Roboter soll die durch das erkannte Thema gewonnene Information nutzbar machen können.

Die Forderung nach Dynamizität schließt in normalen Kommunikationsumgebungen überwachtes Lernen von Themen aus: Eine Phase des überwachten Lernens vor Einsatz des Roboters – also durch den Hersteller – kann aufgrund der Forderung nach Dynamizität nicht oder nur stark eingeschränkt stattfinden. Das explizite Lehren von thematischen Zusammenhängen ist eher selten Teil einer natürlichen Kommunikation, kann also nicht vom Besitzer des Roboters verlangt werden, womit auch die Alternative des überwachten Lernens von Themen nach der Fertigstellung des Robotersystems nicht praktikabel ist[1].

Anhand dieser Anforderung und der kurzen Darstellung der Kernaufgaben von Themenerkennungssystemen im vorigen Abschnitt lässt sich der Vorgang der Themenerkennung auf einem natürlichsprachlich kommunizierenden Robotersystem klassifizieren. Das Robotersystem soll anhand vergangener Kommunikationsvorgänge (unüberwacht) lernen, zu welchem Thema die aktuelle Kommunikation gehört und diese Information möglichst zeitnah den jeweiligen roboterinternen Modulen zur Verfügung stellen. Die Fragestellung, wann genau eine Benutzeräußerung ein neues Thema einleitet (FSD), ist nur implizit relevant, genau wie die Fragestellung, welche Kommunikationsabschnitte dasselbe Thema besitzen (*story link detection*). Es ist viel wichtiger, anhand vergangener Kommunikationsabschnitte etwas über das aktuelle Thema zu lernen und dieses Wissen dem System zur Verfügung zu stellen, als alle Kommunikationsabschnitte desselben Themas aus dem Speicher zu holen. Die Aufgabe eines auf einem Robotersystem arbeitenden Themenerkenners ist also am ehesten als die Aufgabe der *cluster detection* mit einem impliziten Tracking zu klassifizieren. Hinzu kommt die Aufgabe der Extraktion relevanter Informationen über die Themencluster, die den Nachfolgeprozessen vermittelt werden. Diese Aufgabe wird im Kontext dieser Arbeit allerdings als Nebenaufgabe des *cluster detection*-Prozesses angesehen. Eine Ausnahme hierfür bildet jedoch die Erkennung impliziter und expliziter Themennamen, die im Rahmen des *online*-Systems in Kapitel 6 dargestellt wird.

Der Rest des Kapitels beschäftigt sich mit der konkreten Ausarbeitung des Themenerkennungssystems. Um dieser Aufgabe nachgehen zu können, ist aber zuerst eine Definition grundlegender Begriffe nötig.

[1] Tatsächlich gilt dies nur für explizites Lehren von Themen als einzige Quelle thematischer Information. Im Rahmen dieser Arbeit werden wie in Kapitel 6 beschrieben die in natürliche Kommunikation eingebetteten expliziten Themenagaben durchaus zum Training des Themenerkennungssystems verwendet.

3.2.1 Gegenstandsbereich – Begriffsdefinitionen

Es stellt sich die Frage, welche Entsprechung in dialogischer Kommunikation die im TDT-Projekt beschriebenen „Nachrichten“ (*stories*) – also die zu klassifizierenden Entitäten – haben. Es ist trivial zu sehen, dass zeitliche Abschnitte von Kommunikationsereignissen klassifiziert werden sollen, also so genannte *Kommunikationssegmente.*

Kommunikationssegmente seien wie folgt vorläufig definiert:

Definition 3.1 (Kommunikationssegment, unimodal) *Ein* ***unimodales Kommunikationssegment*** *besteht aus allen im Rahmen einer Kommunikation von einem Kommunikationsteilnehmer geäußerten Wortzeichen, die zwischen zwei definierten Zeitpunkten t_1 und t_2 geäußert wurden.*

Bei dieser Definition handelt es sich um eine Arbeitsdefinition, keinen Versuch einer globalen Definition des Begriffs „unimodales Kommunikationssegment“. Aus diesem Umstand erklärt sich die z.B. die Irrelevanz der Reihenfolge der Wörter als auch der Umstand, dass eine Beschränkung auf Wörter vorgenommen wird; alternativ wären auch andere Modalitäten als die verbale Ebene einzubeziehen – die für sich genommen selbst eine Abstraktion gesprochener Sprache ist.

Damit ein Kommunikationssegment eindeutig einem Cluster (Thema) zugeordnet werden kann, gilt weiterhin:

Satz 3.1 *Kommunikationssegmente haben (bezüglich eines Klassifikationsprozesses) immer genau ein oder kein Thema.*

Kommunikationssegmente müssen in den meisten Fällen durch eine *story segmentation* aus dem kontinuierlichen Datenstrom, der eine Kommunikation repräsentiert, gewonnen werden.

Satz 3.1 ist eine vereinfachende Annahme, bei der unklar ist, ob sie in einem konkreten Themenerkennungssystem stets erfüllt ist. Nicht erfüllbar wäre sie, wenn grundsätzlich mehrere Themen bei einem beliebig kleinen Segment, also an einem konkreten Punkt einer Kommunikation, vorliegen können. Trotzdem wird im Rahmen dieser Arbeit davon ausgegangen, dass eine (dialogische) Kommunikation stets in Kommunikationssegmente, für die der obige Satz gilt, unterteilt werden kann. Üblicherweise geschieht dies durch eine genügend geringe Granularität der Segmentierung.

Es stellt sich die Frage, auf Grund welcher Information ein Roboter in der Lage sein soll, Kommunikationssegmente in separate thematische Cluster einzugliedern. Aufgrund der Forderung nach unüberwachtem Lernen kann dies nur anhand von Eigenschaften der Segmente selbst geschehen. (Unimodale) Kommunikationssegmente sind durch die in ihnen vorkommenden Worte bestimmt[2]. Folglich muss diese Information die Grundlage für eine thematische Klassifikation derselben sein. Ein einfaches Themenerkennungssystem würde also Segmente anhand ihrer Ähnlichkeit zu anderen Kommunikationssegmenten in verschiedene Cluster einsortieren. Eine wichtige Vorverarbeitung ist dabei wie beschrieben die Bildung von Kommunikationssegmenten,

[2]Zu beachten ist, dass in der Definition so wie an dieser Stelle von Wort*token*, also Vorkommnissen die Rede ist. Ein Kommunikationssegment enthält somit auch die Information, wie oft ein bestimmter Worttyp in einem Kommunikationssegment instanziiert wurde. Umgangssprachlich: Wie oft ein bestimmtes Wort vorkam.

die nur ein Thema besitzen (*story segmentation*). Diese Vorgehensweise entspricht im Wesentlichen der Vorgehensweise in dieser Arbeit, allerdings findet sich ein wichtiger Unterschied:

Für den Prozess der *cluster detection* besitzen große Kommunikationssegmente mehr Information als kleine, es ist also wünschenswert, möglichst große Segmente zu bilden. Die Anforderung der Echtzeitverarbeitung verbietet jedoch die Klassifikation zu großer Kommunikationssegmente, da das aktuelle Thema erst nach dem Clustern des aktuellen Segments erkannt werden kann. Die in dieser Arbeit gewählte Lösung des Problems besteht in der Differenzierung von zwei Arten von Kommunikationssegmenten:

1. Kurze Segmente, die zur Erkennung des aktuellen Themas herangezogen werden (im Folgenden als „*chunk*“ oder (Benutzer-)Äußerung bezeichnet)

2. Lange Segmente, die als Informationsgrundlage für die *cluster detection* verwendet werden.

Um Ambiguitäten bei der Bezeichnung vorzubeugen, werden im Folgenden nur die unter Punkt 2. aufgeführten Segmente als „Kommunikationssegmente“ bezeichnet. Abbildung 3.1 auf der nächsten Seite dient der Veranschaulichung des bisherigen Aufbaus. *Cluster detection*-Verfahren gliedern sich in zwei Subtypen, nämlich lokale und globale Verfahren. Im folgenden Abschnitt werden die beiden Ansätze dargestellt und gegeneinander abgewogen.

3.2.2 Globale vs. lokale cluster detection-Verfahren

In dem Rahmen des TDT-Projektes wurde zwischen so genannten *lokalen* und *globalen* Ansätzen zur *cluster detection* unterschieden (Allan, 2002b). Globale Ansätze beziehen in ihre Analysen die komplette Datenbasis ein und clustern diese nach Themen. Lokale Ansätze vergleichen eine eingehende, neue Datenstruktur[3] mit schon existierenden Gruppen bekannter Daten und ordnen sie einer dieser (thematischen) Kategorien zu. Auf diese Weise wachsen die Gruppen inkrementell.

Wichtiger Bestandteil der lokalen Algorithmen ist – wie oben beschrieben – die *first story detection* (FSD). Die Aufgabe der FSD ist es zu erkennen, ob eine eingehende Datenstruktur in eine neue, bisher unbekannte Klasse eingefügt werden muss, oder ob sie einer alten Klasse hinzugefügt wird. In globalen Ansätzen ist dies nur implizit der Fall; für alle vorhandenen Daten wird über eine Partitionierung entschieden.

Das TDT-Projekt definierte die Aufgabe der *cluster detection* als nur mit lokalen Algorithmen lösbar. Die Gründe dafür liegen vermutlich in der Komplexität der Algorithmen: Lokale Verfahren können inkrementell arbeiten, und auf diese Weise Ergebnisse bisheriger Berechnungen mit in die aktuelle Berechnung aufnehmen, was zu einer Reduzierung der Rechenlast führt. Globale Verfahren dagegen führen unter Umständen alle Berechnungen neu aus und benötigen so mehr Rechenzeit. Da *cluster detection* idealerweise in Echtzeit ausgeführt wird – die Klassifikation der Datenstruktur als zu einem Thema gehörig soll möglichst kurz nach Eingang der

[3]Im TDT-Projekt also z.B. eine Radionachricht, im Bereich Themenerkennung für Dialogsysteme eine Äußerung.

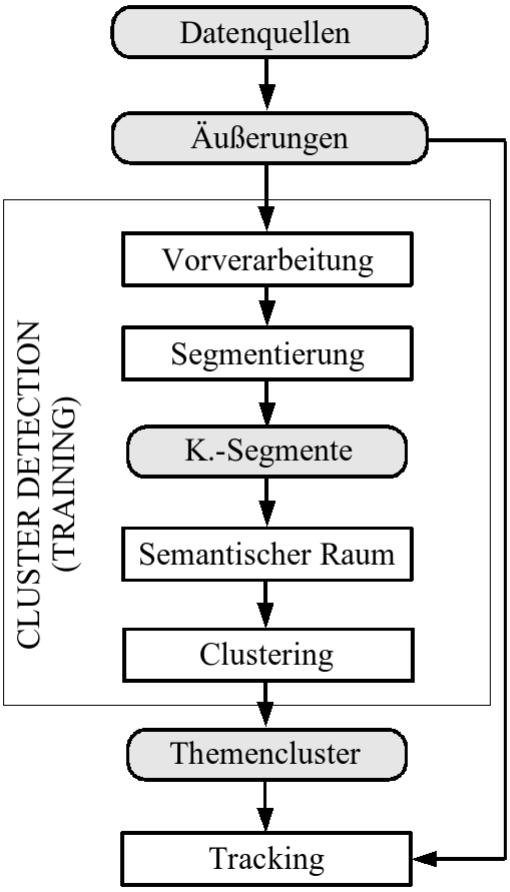

Abbildung 3.1: Vorläufige Basisstruktur der Themenerkennung

Datenstruktur erfolgen – wurden im Rahmen des TDT-Projektes lokale Algorithmen bevorzugt. Letztendlich sind die Anforderungen an die beiden Typen von *cluster detection*-Algorithmen nahezu identisch, der wesentliche Unterschied liegt in der Komplexität, die durch die Vorgehensweise bedingt ist. Da an globale Algorithmen weniger harte Anforderungen gestellt werden, existiert in diesem Kontext auch eine größere Zahl von Ansätzen.

Wichtig ist jedoch, dass trotz der potentiell großen Unterschiede im Laufzeitverhalten die Unterscheidung global vs. lokal nicht mit der Unterscheidung *online*-fähig vs. nicht *online*-fähig zu verwechseln ist. Zwar sind globale Algorithmen weniger gut für Echtzeitberechnungen – und damit *online*-Berechnungen – geeignet, im Rahmen dieser Arbeit wurden sie aber trotzdem in *online*-Kontexten für quasi-Echtzeitberechnungen verwendet (siehe Kapitel 6).

In der folgenden Tabelle 3.1 habe ich noch einmal die Eigenschaften von lokalen und globalen *cluster detection*-Verfahren aufgeführt. Je nach Verfahren können diese Eigenschaften im Detail abweichen.

Lokale *cluster detection*	**Globale *cluster detection***
Inkrementelle Verarbeitung möglich	Keine inkrementelle Verarbeitung möglich
Schnelle Verarbeitung großer Datenmengen	Schnelle Verarbeitung nur auf gesamten Datensätzen
Aufgrund stärkerer Anforderungen nur wenige Algorithmen bekannt	Viele Algorithmen aus dem Bereich des *information retrieval*

Tabelle 3.1: Lokale vs. globale *cluster detection*-Verfahren im Vergleich

In der vorliegenden Arbeit wurde der Fokus dennoch auf die Untersuchung globaler Verfahren zur *cluster detection*, besser Themenerkennung, gelegt. Der Hauptgrund für diese Vorgehensweise liegt in der potentiell größeren Leistungsfähigkeit globaler Verfahren. In den folgenden Abschnitten werden dieser und weitere Gründe für den gewählten Ansatz im Detail diskutiert.

Komplexität Der primäre Grund, aus dem im Rahmen des TDT-Projektes auf globale Verfahren verzichtet wurde, bestand in der Komplexität der relevanten globalen bzw. lokalen Algorithmen. Das Szenario, welches ausschlaggebend für ein *cluster detection*-System im Rahmen des TDT-Projektes ist, besteht z.B. in der Idee eines intelligenten Autoradios, welches auf etlichen Kanälen alle Nachrichtensendungen über mehrere Tage hinweg auf interessante Nachrichten hin analysiert. Die dabei anfallenden Datenmengen lassen nur Algorithmen zu, die eine geringe Komplexität besitzen. Sollte die Komplexität zu groß sein, lässt sich die Anforderung der Echtzeitverarbeitung nicht mehr aufrecht erhalten.

Auch im Kontext einer Themenerkennung situierter Dialogsysteme gilt die Echtzeitanforderung. Allerdings sind die zu verarbeitenden Datenmengen üblicherweise viel geringer: Während im Rahmen des TDT-Projektes normalerweise mit Korpora aus mehreren tausend Nachrichten gearbeitet wurde, besteht z.B. das für diese Arbeit zur Evaluation herangezogene Korpus aus 29 Monologen von insgesamt ca. 300 Minuten Dauer, was aufgrund der großen Anzahl von Pausen zu einer erheblich geringeren Datenbasis führt. In den entwickelten Prototypensystemen existiert

weiterhin ein die Trainingsdatenmenge begrenzender „**Vergessensprozess**“. Abgesehen von dem positiven Seiteneffekt der Einschränkung der Rechenlast bei der *cluster detection* hilft dieser Prozess, mittlerweile irrelevant gewordene thematische Zusammenhänge zu löschen, da diese potentiell zu Fehlklassifikationen führen können. Dies wird durch eine Spezifikation der Maximalgröße der Trainingsdatenbank und eine Definition dieser Datenbank als *queue* erreicht, so dass die ältesten Trainingsdaten mit der Zeit gelöscht werden. Zu beachten ist, das auf diese Weise der auf Seite 8 erwähnte temporale Aspekt von Dynamik – die Veränderung von Themen über Zeit – umgesetzt wird. Nicht zuletzt aufgrund der geringen Größe des jeweiligen Trainingsmaterials wurde er im Verlauf dieser Arbeit allerdings nicht untersucht (die Datenbank erreichte in keinem Fall die maximale Größe) und wird somit nicht weiter diskutiert werden.

Die Annahme, dass die Verarbeitungsgeschwindigkeit von globalen Algorithmen zur *cluster detection* sich im Kontext dieser geringeren Datenmengen[4] in einem echtzeittauglichen Rahmen befindet, ist leider falsch, auch wenn zukünftige schnellere Rechner eine Echtzeitberechnung in mittlerer Zukunft greifbar machen könnten. Allerdings erlaubt eine in dem entwickelten System durchgeführte Trennung von Tracking und *cluster detection* die Klassifizierung einer Datenstruktur als zu einem Cluster gehörig, auch wenn sie auf nicht völlig aktuellen *cluster detection*-Daten beruht. Die *cluster detection* kann somit als zweistufiger Prozess betrachtet werden:

1. Eine globale und kostspielige *cluster detection*, die die Datenbank bekannter Datenstrukturen klassifiziert und strukturiert.

2. Ein in Echtzeit ablaufender Prozess, der neue, eingehende Datenstrukturen anhand der möglicherweise nicht ganz aktuellen global erzeugten Datenbank klassifiziert.

In den durchgeführten Experimenten zeigte sich, dass die Laufzeit der globalen Klassifikation bei mehreren tausend Äußerungen nur wenige Minuten betrug. Da im Allgemeinen davon auszugehen ist, dass in einem solchen Zeitraum während einer situierten Mensch-Roboter-Kommunikation nur wenige Äußerungsereignisse stattfinden, ist die Datenbank normalerweise auf einem aktuellen Stand.

Es bestand im Kontext dieser Arbeit somit kein unabhängiger Grund mehr, aus dem lokale *cluster detection*-Algorithmen globalen Algorithmen vorzuziehen waren. Ich möchte im Folgenden die Gründe anführen, die für eine Verwendung globaler Algorithmen in dieser Arbeit ausschlaggebend sind.

Effektivität Globale Algorithmen wurden im Kontext jahrelanger *information retrieval* und *indexing*-Forschungen entwickelt. Es existiert ein breites Spektrum an Verfahren, die sorgfältig evaluiert wurden. Im Gegensatz dazu sind nur wenige Verfahren zur lokalen *cluster detection* bekannt – selbst im Rahmen des TDT-Projektes wurde dieses Forschungsgebiet eher stiefmütterlich behandelt. Es ist also davon auszugehen, dass hinsichtlich der Effektivität die bekannten globalen Verfahren vorzuziehen sind, zumal sie dieselbe Aufgabe unter geringeren Anforderungen erfüllen.

[4] An dieser Stelle wird auf das Korpus Bezug genommen. Während der *online*-Experimente am Ende dieser Arbeit wurden Echtzeitberechnungen (im Sinne einer Themenerkennung vor der nächsten Benutzeräußerung) durchgeführt.

Die im Rahmen dieser Arbeit praktizierte Bevorzugung leistungsfähiger Algorithmen gegenüber schnellerer, aber weniger effektiver, lässt sich durch die zu erwartenden Schwierigkeiten aufgrund kleiner Trainingsdatenmengen erklären.

Ein weiteres Effektivitätskriterium liegt in dem Anwendungsspektrum der jeweiligen Verfahren: An lokale Verfahren werden höhere Ansprüche bezüglich der Verarbeitungsgeschwindigkeit gestellt. Es ist davon auszugehen, dass die langsameren, globalen Verfahren bessere Ergebnisse erzielen, wenn auf schnelle Berechnung verzichtet werden kann.

Ein Problem bei dem vorgestellten Ansatz ist natürlich die oben genannte Trennung von Tracking und *cluster detection*. Selbst wenn die globale *cluster detection* bessere Ergebnisse liefert, als eine lokale, in Echtzeit ablaufende *cluster detection*, ist der Trackingprozess selbst natürlich auch anfällig für Fehler. Die Gesamtfehlerzahl ist – metaphorisch gesprochen – die Summe der Fehler der beiden Verfahren.

Fehleranfälligkeit der first story detection Ein Charakteristikum der Daten, die in situierten Mensch-Maschine-Dialogen vorkommen, ist die extreme Kürze der zu klassifizierenden Entitäten. Während in dem TDT-Projekt üblicherweise Nachrichtendaten analysiert wurden, bei denen eine gewisse Grundlänge im Normalfall gegeben ist, können von Menschen produzierte Äußerungen extrem kurz sein. Dies führt nach meiner Ansicht zu Problemen bei der *first story detection* als einem wesentlichen Element der lokalen *cluster detection*. Aufgrund des üblicherweise inkrementellen Aufbaus von lokalen Algorithmen zur *cluster detection* können so Fehler entstehen, die nicht mehr revidiert werden können. Dies werde ich im Folgenden genauer erläutern.

Als Beispiel für einen lokalen Algorithmus zur FSD möchte ich (Allan u. a., 1998, S.3) anführen. Der Algorithmus – eine Variante des *single pass*-Algorithmus (vgl. (van Rijsbergen, 1979, S.35f)) gliedert sich wie folgt:

1. Verwende Verfahren zur Merkmalsextraktion und Selektionstechniken, um eine Anfragerepräsentation des Inhalts des zu klassifizierenden Dokuments aufzubauen.

2. Bestimme den initialen Schwellwert des Dokuments, indem das Dokument mit der neu gewonnenen Anfragerepräsentation evaluiert wird.

3. Vergleiche das Dokument mit den früheren Anfragerepräsentationen im Speicher.

4. Wenn das Dokument zu keiner alten Anfragerepräsentation passt – also den gemessenen Schwellwert bei keiner alten Anfragerepräsentation überschreitet – markiere das Dokument als neu.

5. Wenn das Dokument zu einer alten Anfragerepräsentation passt, markiere das Dokument als nicht neu.

6. (Optional) Füge die Anfragerepräsentation zu der Liste der aufgelösten Anfragerepräsentationen hinzu.

7. (Optional) Strukturiere die existieren Anfragerepräsentationen anhand des neuen Dokuments um.

8. Speichere die neue Anfragerepräsentation.

Der große Vorteil dieses Algorithmus liegt eindeutig in dem Umstand, dass seine Verarbeitungsgeschwindigkeit nur linear von dem Produkt der Anzahlen der Dokumente und Themen im Speicher abhängig ist[5]. Allerdings verfügt der Algorithmus über keinen Reparaturmechanismus für begangene Fehler der FSD, die durch neu erworbenes Wissen korrigiert werden könnten. Dazu ein (konstruiertes) Beispiel aus dem Bereich der situierten Mensch-Maschine-Kommunikation. Gesetzt den Fall, eine Person instruiert einen Roboter wie folgt:

Beispiel 3.1 *this is a red ball*

Beispiel 3.2 *there is a brown table*

Der oben genannte Algorithmus würde die beiden Äußerungen mit sehr hoher Wahrscheinlichkeit zu unterschiedlichen Themen zuordnen, da die Nicht-Funktionswörter völlig unterschiedlich sind. Nur eine Trainingsbasis, in der die Nicht-Funktionswörter schon miteinander assoziiert wurden, könnte dies unterbinden. Auf jeden Fall kann es aber bei mangelndem Trainingsmaterial geschehen, dass (fälschlicherweise) für die zweite Äußerung eine neue Themengruppe eingeführt wird. Selbst wenn die Einsortierung in Themen nicht anhand einer Anfragerepräsentation, sondern anhand von multimodalen Merkmalen geschieht, können selbstverständlich Fehler auftreten.

Angenommen, für Beispiel 3.2 wurde ein neues Thema eingeführt. Anschließend wird der Satz aus Beispiel 3.3 geäußert:

Beispiel 3.3 *the red ball lies on the brown table*

Diese Äußerung markiert Beispiel 3.1 und Beispiel 3.2 als zu einem gemeinsamen Thema zugehörig. Der genannte Algorithmus ist aber nun nicht mehr in der Lage, die Themen nachträglich zu verschmelzen. Ebenso wäre er nicht in der Lage, bei einer fehlerhaften Zuordnung von zwei Äußerungen zu einem gemeinsamen Thema die Äußerungen wieder nachträglich zu trennen, wenn nachfolgende Äußerungen dies nahelegen würden.
Natürlich sind Algorithmen denkbar, die diese Probleme lösen. So wäre z.B. eine Modifikation des *single pass*-Algorithmus möglich, die Themen miteinander verschmilzt, wenn mehrere Äußerungen zu beiden Themen zugeordnet werden würden. Eine Lösung des konträren Problems wäre dagegen schwieriger. Auf jeden Fall stellt sich die Frage, ob solche Fehlerkorrekturmechanismen nicht die Echtzeiteigenschaften von lokalen *cluster detection*-Algorithmen soweit begrenzen, dass die in dieser Arbeit durchgeführte Teilung der Aufgabe in Tracking und globale *cluster detection* zumindest keinen Nachteil in der Komplexität gegenüber lokalen Algorithmen besitzt.

Modularität Ein letztes Argument gegen die Verwendung von lokalen Algorithmen zur *cluster detection* beruht auf den besseren Strukturierungsmöglichkeiten von globalen Algorithmen. Der beschriebene FSD-Algorithmus ist – nach Einbeziehung der optionalen Punkte 6. und 7. – sowohl ein FSD-, als auch ein Cluster- und ein *cluster detection*-Algorithmus so wie ein unüberwacht arbeitender Trackingalgorithmus. In

[5]Unter der vereinfachenden Annahme, dass der Vergleich einer Anfragerepräsentation mit einer anderen in konstanter Zeit stattfindet und nur eine solche je Thema existiert.

dem von mir entwickelten Verfahren werden diese Aufgaben von unterschiedlichen Modulen und Algorithmen übernommen, Ausnahme ist dabei nur die potentiell fehleranfällige FSD, die implizit in der *cluster detection* enthalten ist. Auf diese Weise kann jeder einzelne Schritt optimiert und gegen Alternativverfahren ausgetauscht werden, was einen Vorteil gegenüber dem vorgeschlagenen Algorithmus darstellt.

Allans Algorithmus hat natürlich ein anderes Aufgabengebiet als das in dieser Arbeit behandelte. Auch habe ich in den vergangenen Abschnitten nur exemplarisch Argumente vorgebracht, die nicht zwangsläufig auf alle lokalen Verfahren zur *cluster detection* anwendbar sind. Trotzdem glaube ich in den letzten Abschnitten plausibel gemacht zu haben, warum ein globaler Ansatz nicht zwangsläufig nachteilig für das Themenerkennungssystem ist.

3.2.3 Zusammenfassung

In dem vorliegenden Kapitel wurden verschiedene Gründe für die Verwendung globaler *cluster detection*-Verfahren erwogen. Zusammenfassend lässt sich sagen, dass in dieser Arbeit eine mögliche Einschränkung der Echtzeitfähigkeiten der Themenerkennung – oder zumindest eine Einschränkung in der Echtzeit-Eingliederung neuer Information in die Trainingsdatenbank – hingenommen wird, um den Problematiken von lokalen Verfahren, insbesondere der fehlenden Fehlerkorrektur vorzubeugen. Die hingenommene Einschränkung sollte sich aufgrund der geringen zu erwartenden Trainingsdatenmengen als unwesentlich erweisen. Weiterhin versprechen globale Algorithmen bessere Resultate, die angesichts der kleinen Trainingsdatenmengen notwendig sind.

Die Entscheidung für ein globales *cluster detection*-Verfahren in Kombination mit einem separaten Trackingvorgang legt die grundlegende Struktur des Systems fest. Es stellt sich jetzt die Frage, wie diese beiden Teilaufgaben bewältigt werden können und welche weiteren Schritte zur Unterstützung und Verbesserung der Ergebnisse benötigt werden.

Im Rahmen dieser Arbeit wurden verschiedene Evaluationen des Systems durchgeführt, die sich insbesondere in Bezug auf die Aufgaben der Vorverarbeitung und der Segmentierung unterscheiden. Für die Aufgaben der Themenerkennung und Klassifizierung (Tracking) wurden im Rahmen der Evaluationen jedoch stets dieselben Methoden verwendet. Aus diesem Grund möchte ich im Folgenden die gewählten Ansätze, mögliche Alternativen und die Gründe für die jeweilige Wahl darstellen.

3.3 Vorüberlegungen zur Klassifikation

Üblicherweise werden in globalen *cluster detection*-Ansätzen Dokumente (Nachrichten, Äußerungen...) in Gruppen einsortiert, d.h. klassifiziert. Um wie geschildert Quasi-Echtzeitfähigkeiten des Themenerkennungssystems zu ermöglichen, werden neu eintreffende *chunks* anhand der gebildeten Themeninformation klassifiziert. Eine Zuordnung neu eintreffender *chunks* zu einem bekannten Cluster muss anhand einer Ähnlichkeitsbeziehung zwischen dem Cluster (von Kommunikationssegmenten) und dem *chunk* geschehen: Das Cluster, dessen Elemente – also die Kommunikationssegmente – dem *chunk* am ähnlichsten sind, ist das korrekte.

Für diese Arbeit wurde ein anderer Weg beschritten: Nicht die Kommunikationselemente wurden nach Themengruppen sortiert, sondern die in den Kommunikationselementen vorkommenden relevanten Worte bzw. Zeichen. Die Klassifikation geschieht dann anhand der Vorkommnisse von zu Themen gehörenden Worten in noch nicht klassifizierten *chunks*. Dieser Ansatz birgt zwar den Nachteil, dass auf diese Weise viele gängige Klassifikationsverfahren wie z.B. Support-Vector-Maschinen nicht anwendbar sind (s.u.), die möglicherweise zu besseren Klassifikationsergebnissen führen würden. Der entscheidende Vorteil besteht jedoch in dem direkten Nutzen, den als Zeichengruppen angegebene Themen mit sich bringen: So lassen sich z.B. (in der multimodalen Variante des Verfahrens) Informationen über thematisch-kontextuell relevante Objekte – da sie selbst als Vektoren in einem Cluster repräsentiert sind – direkt inferieren; eine solche Inferenz ist anhand von Clustern von Kommunikationssegmenten auch möglich, jedoch wesentlich aufwändiger.

Prinzipiell möchte ich an dieser Stelle keinen der beiden Wege als Königsweg darstellen. Ob sich mit Hilfe von alternativen Klassifikationsverfahren wirklich relevant bessere Ergebnisse in der Klassifikation erzielen lassen, ist aufgrund der extremen Kürze der *chunks*, die z.B. in dem Experimentkorpus oftmals nur ein einzelnes themenanzeigendes Wort enthalten, fraglich. Wie eine Evaluation auf dem Reuters-21578-Korpus gezeigt hat (vgl. Abschnitt 6.6), können mit dem verwendeten Verfahren auch in üblicheren Anwendungsgebieten Ergebnisse erzielt werden, die zumindest nicht gravierend schlechter sind als mit etablierten Methoden. Mögliche zukünftige Untersuchungen bezüglich der Themenerkennung auf Robotersystemen sollten versuchen, den optimalen Weg zu finden.

Die folgenden beiden Unterkapitel beschäftigen sich mit den ausgewählten Verfahren, die zur *cluster detection*, also zur Bildung von Themengruppen von Worten, gewählt wurden. Im Wesentlichen sind dabei zwei Teilschritte zu differenzieren: Die Berechnung von Distanzen von Worten und das Clustern der Worte in Themengruppen. Im anschließenden Unterkapitel wird dann genauer auf den Prozess der Klassifikation eingegangen. Der Abschluss dieses Kapitels widmet sich Überlegungen, wie die Effektivität des Verfahrens mit bekannten so wie neuen Methoden verbessert werden kann.

3.4 Semantische Räume

> [Es] [...] interessieren [...] ja nicht die individuellen Beiträge der daran beteiligten Einzelnen, sondern vielmehr die erst durch (mehr oder weniger) gleichartigen Gebrauch sprachlicher Terme von Vielen [...] sichtbar werdenden *Verwendungsregularitäten* [...]
> *B. Rieger*[6]

Ein wichtiger Teilaspekt einer Clusterung – die als Grundlage der *cluster detection* angesehen werden kann – besteht in der Berechnung von (thematischen) Distanzen zwischen den zu klassifizierenden Entitäten. In der Linguistik existiert seit geraumer Zeit der Begriff „semantischer Raum“ als Bezeichnung für eine Projektion von linguistischen Entitäten in einen Raum, der die semantischen (und thematischen) Eigenschaften der Entitäten durch Distanzen zwischen denselben widerspiegelt.

[6](Rieger, 1989, S.196), Löschungen und Einfügungen in Klammern von mir.

Diese Formulierung ist bewusst schwach gewählt. So definiert Lowe:

Definition 3.2 (semantic space (Lowe)) *A semantic space model is [a] method of assigning each word in a language to a point in a real finite vector space.*[7]

Auch wenn diese Definition einen guten Eindruck davon vermittelt, was einen semantischen Raum ausmacht, bindet sie den Begriff des semantischen Raumes an den des Vektorraumes, was meiner Ansicht nach eine unnötige Einschränkung ist – Beispiele für nicht-vektorraumbasierte semantische Räume werden am Ende dieses Abschnitts diskutiert, auch wenn sie im Rahmen dieser Arbeit keine Verwendung finden. Außerdem schränkt Lowe mit dieser Definition semantische Räume auf Räume von *Worten* ein – eine Einschränkung, die im Rahmen dieser Arbeit als unnötig widerlegt wird.

Eine wesentlich allgemeinere Definition findet sich in (Leopold, 2005):

Definition 3.3 (semiotic system) *A semiotic system $\tilde{S}$ consists of signs s. Signs fulfil a communicative function $f(s)$ within the semiotic system to meet the communicative requirements of [the] system's user.*[8]

Definition 3.4 (semantic space (Leopold)) *Let $\tilde{S}$ be a semiotic system, (S,d) a metric space and r: $\tilde{S} \rightarrow S$ a mapping from $\tilde{S}$ to S. A Semantic Space (S,d) is a metric space whose elements are representations of signs of a semiotic system, i.e. for each $x \in S$ there us a $s \in \tilde{S}$ such that r(s) = x. The inverse metric $(d(x,y))^{-1}$ quantifies some functional similarity of the signs $r^{-1}(x)$ and $r^{-1}(y)$ in S.*[9]

Die Forderung nach einem Vektorraum entfällt in dieser Definition. Außerdem sind semantische Räume nicht mehr auf Worte festgelegt – jedes semiotische System beliebiger Modalität kann in einen semantischen Raum projiziert werden.

Semantische Räume wurden in den Anfängen manuell konstruiert, mittlerweile ist ihre Erzeugung aus Korpora üblich (Lund und Burgess, 1996). Wesentliches Indiz für die Berechnung von Zeichen repräsentierenden Punkten ist dabei lexikalische Kookkurrenz in (durch die Datenlage bzw. passende Vorverarbeitungsprozesse festgelegten) Kontexten. Die Lage eines Punktes ist also definiert durch die Kookkurrenz-Verteilung aller in Beziehung gesetzten Zeichen. Solche semantischen Räume verfügen also über keine vordefinierten Raumdimensionen (wie es z.B. Themengebiete sein könnten, denen die Zeichen mehr oder weniger entsprechen), sondern bilden die Dimensionen anhand der zugrundeliegenden Datenlage.

Betrachtet man semantische Räume, die aus Korpora aufgebaut wurden und die sich lexikalische Kookkurrenz zunutze machen, so stellt man folgendes fest: Die reine

[7] (Lowe, 2001, S.2), Worte in eckigen Klammern von mir hinzugefügt.

[8] (Leopold, 2005, S.1), Worte in eckigen Klammern von mir hinzugefügt. Durch den Bezug auf die „*communicative requirements*“ stellt sich Leopold in die Tradition von Zipf (Zipf, 1949), Altmann (Altmann, 1981) und Köhler (Köhler, 1986). Zipf formulierte zwei Bedürfnisse von Sprachbenutzern (Minimierung des Gedächtnisaufwands und Minimierung des Produktionsaufwands), denen Sprache als kommunikatives System mehr oder weniger gerecht werden kann. Altmann (Altmann, 1981) formuliert den Gedanken von Sprache als einem selbstregulierendem System. Köhler erweiterte diese Ansichten und postulierte die Existenz von Sytembedürfnissen. Bei diesen handelt es sich um Elemente der (Sprach-)Systemumgebung, die aufgrund ihrer Relation zu den Elementen des Systems Änderungen in demselben hervorrufen können (Köhler, 1986, S.43).

[9] (Leopold, 2005)

Untersuchung von Kookkurrenz zweier Symbole in einem Kontext, bzw. die Assoziation von semantischer Ähnlichkeit in einem Korpus mit einem Maß des Miteinander-Vorkommens innerhalb des Korpus erlaubt nicht die Analyse von *paradigmatischer* Ähnlichkeit. Dazu ein Beispiel: Die Worte „Tanne“ und „Fichte“ bezeichnen in einem (fiktiven) Korpus zwei Baumarten, die semantisch sehr nah miteinander verwandt sind. Ggf. fehlt den Quellen des Korpus sogar die Unterscheidungsfähigkeit zwischen den beiden Baumarten, so dass sie – in diesem Korpus – als Synonyme angesehen werden können. Betrachtet man aber – zumindest auf Satzniveau – die Kookkurrenz dieser beiden Worte, so kann sie leicht gegen null gehen. Dieser Effekt erklärt sich dadurch, dass immer, wenn das Wort „Fichte“ verwendet wurde, die Verwendung des Wortes „Tanne“ nicht mehr nötig ist. Es handelt sich folglich um komplementär verteilte „Synonyme“. Diese semantische Ähnlichkeit würden „semantische“ Räume, die ausschließlich auf direkten Kookkurrenzinformationen basieren, nicht berücksichtigen. Sehr wohl würden sie aber Aspekte der *syntagmatischen* Beziehungen z.B. zwischen „Nadeln“ und „Tanne“ messen können – in diesem Fall ihr (vermutlich) häufig gekoppeltes Auftreten.

Diese Erkenntnis floss mit ein in die Entwicklung der Fuzzy Semantics – dem semantischen Raum nach Rieger; dieser spezielle semantische Raum wird in Abschnitt 3.4.2 genauer beschrieben. Eine zentrale Idee dieses Raumes besteht in einem zweischrittigen Prozess, bei dem zuerst die syntagmatischen Beziehungen in einem Raum – dem **Korpusraum** – abgebildet werden, um dann aber in einem zweiten Schritt die paradigmatischen Beziehungen ebenfalls in einen Raum – den **Bedeutungsraum** – projiziert messen zu können (Rieger, 1989) (Mehler, 2001).

Im Kontext dieser Tradition findet sich in (Mehler, 2006) eine Definition des Begriffs „semantischer Raum“:

Definition 3.5 (preliminaries) *Let* $C = \{x_1, \ldots, x_n\}$ *be a text corpus,* $\mathbb{S}$ *a segmentation mapping each text* $x \in C$ *onto an ordered rooted tree* $\mathbb{S}(x) = (S(x), E, x, O_1, O_2)$ *as a model of its kernel hierarchical structure in the sense of an* ordered hierarchy of content objects $[\ldots]$ *and* $\mathbb{L} : T(C) \to L(C)$ *a lemmatization mapping each token* $\boldsymbol{a} \in T(C)$ *onto its type* $a \in L(C); T(C) \subset S(C)$ *is the set of tokens and* $L(C)$ *the set of types of corpus* C*.* O_1 *is an order relation maping the syntagmatic order of all immediate constituents of any segment of* x*. That is,* $O_1(y_i, y_j)$ *iff* $y_i, y_j \in S(x)$ *are immediate constituents of the same* $z \in S(x)$ *according to* $\mathbb{S}$ *so that* y_i *precedes* y_j *in* z*.* O_2 *is the linear order relation induced by the postorder traversal of* $\mathbb{S}(x)$*.*

We define $S(x)$*,* $x \in S(x)$*, as the set of all segments of* x *according to* $\mathbb{S}$ *and* $S(C) = \bigcup_{x \in C} S(x)$*. Further,* $T(x) \subset S(x)$ *is the set of all tokens of* x *according to* $\mathbb{S}$ *and* $T(C) = \bigcup_{x \in C} T(x)$*. Next,* $L(x) = \{a \mid \exists \boldsymbol{a} \in T(x) \models_T a\}$ *is the set of all types classifying at least one token in* $T(x)$*. Thus,* $L(C) = \bigcup_{x \in C} L(x)$*. We write* S, T *and* L *instead of* $S(C), T(C)$ *and* $L(C)$ *if the corpus* C *is known from the context.* $\mathbb{L}$ *induces a* type-token classification $(T, L, \models_T)$ *where* $\models_T \subseteq T \times L$ *and* $\boldsymbol{a} \models_T a$ *iff* $\boldsymbol{a} \in T$ *is a token according to* $\mathbb{S}$ *instantiating the type* $a \in L$ *according to* $\mathbb{L}$ $[\ldots]$[10]

Definition 3.6 (semantic space (Mehler)) *Let a Corpus* C*, a segmentation* $\mathbb{S}$ *and a lemmatization* $\mathbb{L}$ *be given according to definition [3.5]. Further, let* $\mathbb{X}$ *be an uncountable set, e.g.* $\mathbb{X} = \mathbb{R}^n$ *for some* $n{>}0$*,* $n \in \mathbb{N}$*, and* $(\mathbb{X}, d)$ *be a metric space.*

[10] Löschungen in eckigen Klammern von mir.

A semantic space *is a quintuple* $(L, S, \alpha, \beta, (\mathbb{X}, d))$ *where* $\alpha : L \to \mathbb{X}$ *is a function mapping types* $a \in L$ *onto representations of the contexts of their tokens* $\boldsymbol{a} \in L$ *in segments* $x \in S$*. Further,* $\beta : S \to \mathbb{X}$ *is a function mapping segments* $x \in S$ *onto* $\mathbb{X}$ *by operating on the context representations of their components according to* $\mathbb{S}$ *down to the level of tokens* $\boldsymbol{a} \in T(x)$ *as instances of types* $a \in L(x)$[11]

Diese Definition drückt eine etwas striktere Interpretation des Begriffs „semantischer Raum“ aus, als sie in dieser Arbeit vertreten wird. Ein wichtiger Unterschied beruht auf der Definition von C als Textkorpus. Wie weiter oben angedeutet, wird in dieser Arbeit die Position vertreten, dass auch Korpora, die nicht ausschließlich aus Wortsymbolen bestehen, durch semantische Räume abgebildet werden können. Die Einschränkung auf Textkorpora findet sich ebenfalls konnotativ in der festen Eingliederung von einem Lemmatisierungsprozess $\mathbb{L}$, der Symbole auf Symboltypen abbildet. Dieser Prozess ist aber für multimodale Korpora unkritisch, da nicht zwangsläufig eine echte grammatische Lemmatisierung – also eine Abbildung von Worten auf Lemmata – vorliegen muss, wie der Name nahelegt: Die Abbildung nur der Symbolvorkommnisse, die durch dasselbe Symbol repräsentiert sind, auf einen Typen ist ebenfalls möglich und stellt somit eine „Minimallemmatisierung“ dar.

Weiterhin existiert durch die Ordnung der Segmente und Vorkommnisse in einem *ordered rooted tree* die Forderung, dass sich die einzelnen Symbole in einer linearen Ordnung befinden müssen – aber z.B. bei Symbolen, die sprachbegleitend geäußert werden, ist unklar, ob eine solche lineare Ordnung immer vorliegen kann.

Letztendlich stellen diese Differenzen aber nur ein marginales Problem dar, eine Umdefinierung auf Korpora bestehend aus multimodalen Symbolen ist prinzipiell unproblematisch.

Der zentrale Punkt, der diese Definition von den bisherigen Definitionen trennt, ist jedoch die Postulierung der beiden Funktionen α und β. α bildet lediglich die Typen anhand von Repräsentationen, die durch die Verteilung ihrer Vorkommnisse in Segmenten gewonnen wurden, in dem metrischen Raum ab. Ein Beispiel für eine solche Repräsentation wird in Abschnitt 3.4.1 auf der nächsten Seite gegeben. Diese Abbildung gibt Aufschluss über die syntagmatischen Ähnlichkeitsbeziehungen zwischen den Wort-Typen. Für die Darstellung paradigmatischer Ähnlichkeitsbeziehungen müssen erst in einem zweiten Schritt die Beziehungen der Segmente untereinander betrachtet werden, wofür die durch die α-Funktion gewonnenen Typenrepräsentationen ihrer Vorkommnisse (Worte) Verwendung finden. Es handelt sich also um einen im Wesentlichen zweistufigen Prozess, der am detailliertesten in dem semantischen Raum nach Rieger modelliert wurde. Wichtig ist jedoch zu bemerken, dass die Definition nur sehr geringe Anforderungen an die inneren Zusammenhänge der α- und β-Funktionen stellt – über die Modellierung dieser Funktionen definiert sich zu großen Teilen der Typ des semantischen Raumes.

In der Geschichte des *information retrieval* haben sich semantische Räume wie die LSA als geeignetes Mittel erwiesen, um die Probleme von *lexical matching*-Verfahren im Umgang mit Polysemie und Synonymie zumindest teilweise zu kompensieren (Sahlgren, 2001) (Lund und Burgess, 1996) (Deerwester u. a., 1990) (Schütze, 1992). Dies ist zumindest teilweise auf die Fähigkeit semantischer Räume, paradigmatische Beziehungen zwischen Worttypen abzubilden, zurückzuführen.

[11] Referenz in eckigen Klammern von mir angepasst.

In dieser Arbeit wurden verschiedene semantische Räume verwendet, um eine Grundlage für das Clustern nach Themengebieten zu schaffen. Die Gründe dafür sind naheliegend: Diese Arbeit profitiert von den umfangreichen Vorarbeiten, außerdem wurden semantische Räume genau für die Aufgabe der unüberwachten thematisch-semantischen Ordnung von Zeichen aufgrund von Verteilungsinformation erfunden. Weiterhin wird in dieser Arbeit der Versuch unternommen, semantische Räume durch die Bildung von thematisch zusamenhängenden Trainingskontexten – Kommunikationssegmenten – dahingehend beeinflussen, dass sie auf die Repräsentation thematischer Ähnlichkeiten optimiert sind.

Ich möchte im Folgenden verschiedene semantische Räume darstellen, wobei ich mich stark an (Leopold, 2005) orientieren werde. Dabei werde ich zu jedem Raum diskutieren, ob und warum (bzw. warum nicht) er im Kontext dieser Untersuchung verwendet wurde. Es ist zu beachten, dass durch die jeweiligen Räume – und insbesondere die in dieser Arbeit verwendeten – nicht in jedem Fall auf ein Distanzmaß festgelegt sind. Aus diesem Grund werde ich im Anschluss kurz mögliche Distanzmaße in semantischen Räumen diskutieren.

3.4.1 Einfaches Vektorraummodell

Bei dem einfachen Vektorraummodell handelt es sich um eine simple, direkte Projektion von Wortverteilungen in einen Vektorraum. Jede Dimension n_k eines Wortvektors entspricht einem Kontext (Nachrichtenartikel, Dokument o.ä.), die Anzahl der Dimensionen (K) kann folglich sehr groß werden. Die Werte $v_1 \dots v_k$ eines Vektors V für eine bestimmte Dimension entsprechen der relativen oder absoluten Auftretenshäufigkeit[12].

Da die Reihenfolge der Worte in dem jeweiligen Dokument keine Bedeutung für den Aufbau des Vektorraums hat, spricht man von einem *bag of words*-Ansatz. Die zugrundeliegende Information, die der Vektorraum repräsentiert, ist folglich die Kookkurrenzinformation, also die Information, welche Worte wie oft miteinander in Dokumenten vorkommen. Die Erkenntnis, dass aus dieser Information Wissen über thematische und semantische Sachverhalte gewonnen werden kann, stellt die Grundlage vieler semantischer Räume und vergleichbarer *text mining*-Verfahren dar.

Repräsentiert werden kann der einfache Vektorraum durch eine so genannte *term document matrix* (Term-Dokument-Matrix) (Salton und McGill, 1983):

Definition 3.7 (Term-Dokument-Matrix) *Sei C ein Korpus aus Texteinheiten (Dokumenten). W ist die Anzahl der verschiedenen Worte w_1 bis w_W in C. C ist definiert als die Menge der Dokumente d_1 bis d_D, wobei D die Anzahl der Dokumente in C ist. Der Begriff Term-Dokument-Matrix von einem Korpus C ist definiert als*

$$A = (f(w_i, d_j))_{i=1,\dots W, j=1,\dots D} \tag{3.1}$$

wobei f die Häufigkeit (frequency) eines Wortes w_i in dem Dokument d_j darstellt.

Zur Veranschaulichung möchte ich an dieser Stelle ein Beispiel anführen. Gegeben sei ein Beispielkorpus, bestehend aus den folgenden (fiktiven) Titeln (Dokumente) wissenschaftlicher Texte:

[12] Im Folgenden wird von der absoluten Auftretenshäufigkeit ausgegangen.

	D1	D2	D3	D4
EPS	1	0	0	0
graph	0	1	0	1
human	0	0	1	0
interface	1	0	1	0
machine	0	0	1	0
management	1	0	0	0
order	0	0	0	1
sort	0	0	0	1
survey	0	1	0	0
system	1	0	0	0
tree	0	1	0	3

Tabelle 3.2: Term-Dokument-Matrix des Beispielkorpus mit absoluten Häufigkeiten

	D1	D2	D3	D4
EPS	1	0	0	0
graph	0	0.5	0	0.5
human	0	0	1	0
interface	0.5	0	0.5	0
machine	0	0	1	0
management	1	0	0	0
order	0	0	0	1
sort	0	0	0	1
survey	0	1	0	0
system	1	0	0	0
tree	0	0.25	0	0.75

Tabelle 3.3: Term-Dokument-Matrix des Beispielkorpus mit relativen Häufigkeiten

Beispiel 3.4
d1: The EPS interface management system
d2: Graph and trees: a survey
d3: Human machine interfaces
d4: Ordered trees, sorted trees and graph trees

Die dazugehörige, ungewichtete Term-Dokument-Matrix absoluter Häufigkeiten ist in Tabelle 3.2 dargestellt, die ungewichtete Term-Dokument-Matrix relativer Häufigkeiten in Tabelle 3.3. Funktionswörter wurden dabei nicht in die Matritzen aufgenommen, weiterhin wurde eine Lemmatisierung durchgeführt.

Es ist einleuchtend, dass in Bezug auf übliche Korpora meist Matritzen entstehen, die an den meisten Positionen den Wert 0 haben (*sparse matrix*), da viele Worte nur in wenigen Dokumenten vorkommen.

Dem einfachen Vektorraummodell, wie auch den beiden folgenden, ebenfalls auf einem Vektorraum basierenden Verfahren ist gemein, dass theoretisch nicht nur Wortvektoren in einem durch Dokumente aufgespannten Raum, sondern auch Dokumentvektoren in einem durch Worte aufgespannten Raum betrachtet werden können. Beim einfachen Vektorraum geschieht dies einfach durch eine Transposition der Term-

Dokument-Matrix. Auf diese Weise kann man nicht nur Wortähnlichkeiten, sondern auch Dokumentähnlichkeiten messen.

Exkurs: Segmentierung Wie geschildert beruhen alle *bag of words*-basierten Ansätze auf der Ausnutzung von Kookkurrenzinformation. Folglich gilt, dass die Fähigkeit des semantischen Raumes, globale[13] thematische Zusammenhänge zu repräsentieren mit der Qualität der zugrundeliegenden Segmentierung, aber auch mit der Größe der Trainingsmenge eng verknüpft ist. Im Umkehrschluss bedeutet dies, dass geringe Trainingsmengen – wie sie bei multimodaler HRI angenommen werden können – eine sehr gute Segmentierung notwendig machen, während auf sehr großen Korpora z.B. einfach Segmente gebildet werden können, indem nach festen Wortanzahlen eine Grenze eingefügt wird (Rehder u. a., 1998).

Die Segmentierung wurde bereits in Abschnitt 3.2.1 auf Seite 33 als relevanter und eigenständiger Vorverarbeitungsschritt für eine *cluster detection* identifiziert. Sie ist allerdings im Gegensatz zu den semantischen Räumen und Klassifikatoren – die ja letztendlich sehr generische Methoden sind – eng mit der Datengrundlage verknüpft. Deshalb wird auf das Problem der Segmentierung in den Kapiteln eingegangen, in denen die experimentellen Ergebnisse vorgestellt werden.

Im Rahmen dieser Arbeit gilt, dass eine gute Segmentierung einen Text/Monolog/Dialog möglichst exakt und vollständig an den Themengrenzen unterteilt. Auf Themenhierarchien innerhalb eines Textes kann dabei nicht eingegangen werden, daher sollte die Granularität der Segmentierung anhand der von dem Themenerkennungssystem profitierenden Prozessen passend gewählt werden.

Exkurs: Paradigmatische Relationen im einfachen Vektorraummodell Dem aufmerksamen Leser wird nicht entgangen sein, dass das einfache Vektoraummodell dem Standardfall des von Mehler zurückgewiesenen Raumes entspricht, der keine paradigmatischen Ähnlichkeiten zu fassen in der Lage ist. Es existiert keine β-Funktion, mit der ein solcher Schritt vollzogen werden kann. Aus diesem Grund handelt es sich bei dem einfachen Vektorraummodell um keinen semantischen Raum nach (Mehler, 2006). Warum ist das einfache Vektorraummodell in dieser Arbeit also als semantischer Raum präsentiert worden?

Die Antwort auf diese Frage findet sich in den nachfolgenden Arbeitsschritten. Wie an späterer Stelle dargestellt, basiert der in dieser Arbeit verwendete Klassifikationsprozess nicht direkt auf den „Ähnlichkeiten“, die von den semantischen Räumen jeweils angegeben werden, sondern auf Themenclustern, die durch einen Clusterprozess auf den Räumen gewonnen werden. Innerhalb eines solchen Clusters kann es aber sehr wohl geschehen, dass zwei Worte, die selten oder nie miteinander in einem Korpus vorkommen, miteinander assoziiert werden – in dem o.g. Beispiel können „Fichte“ und „Tanne“ durch ihre gemeinsame Nähe zu „Nadeln“ in dasselbe Cluster aufgenommen werden. Somit existiert eine Form von paradigmatischer Assoziation,

[13] D.h. Zusammenhänge, die nicht nur die thematische Struktur der Datengrundlage des semantischen Raumes widerspiegeln, sondern die thematische Struktur aller Dokumente des jeweiligen Anwendungskontexts. In Bezug auf Themenerkennung in situierter Mensch-Roboter-Kommunikation ist dies oft ein temporaler Zusammenhang: Die Datenbasis des semantischen Raums besteht aus den vergangenen Dialogen, die Gesamtheit der Dokumente aus dem Anwendungskontext umfasst zusätzlich die zukünftigen Dialoge.

die über das einfache Vektorraummodell hinausgeht.

Im Licht dieser Betrachtung möchte ich den Leser bitten, die Bezeichnung des einfachen Vektorraummodells als „semantischen Raum“ hinzunehmen, obwohl gegen diese Vorgehensweise zweifellos begründete theoretische Bedenken existieren können.

3.4.2 Semantischer Raum nach Rieger (Fuzzy Semantics)

Eine Modifikation des einfachen Vektorraummodells wurde von Burkhard B. Rieger vorgeschlagen (Rieger, 1989) (Rieger, 1981) (Mehler, 2001). Der Grundgedanke von Riegers Verfahren besteht in der expliziten Trennung der Untersuchung von syntagmatischen und paradigmatischen Verwendungsregularitäten.

Semantische Räume nach Rieger werden somit in zwei getrennten Schritten – der α- und β-Abstraktion erzeugt. Die α-Abstraktion wird auch als *syntagmatische* Abstraktion bezeichnet, die β-Abstraktion als *paradigmatische*.

α-Abstraktion Die α-Abstraktion stellt im Wesentlichen eine modifizierte Berechnung des Korrelationskoeffizienten dar. Sie ist wie folgt definiert:

$$\alpha_{i,j} = \frac{\sum_{k=1}^{D}((g(w_i,d_k))(g(w_j,d_k))}{\sqrt{\sum_{k=1}^{D}((g(w_i,d_k))^2(g(w_j,d_k))^2)}} \in [-1;1] \tag{3.2}$$

wobei

$$g(i,j) = (f(w_i,d_k) - E(f(w_i \mid d_k)) \tag{3.3}$$

E wird anhand einer Abschätzung der Wortverteilungen in allen Dokumenten berechnet:

$$E(f(w_i \mid d_k) = f(w_i)\frac{f(d_k)}{L} \tag{3.4}$$

L ist dabei die Anzahl aller Wortvorkommnisse in C.

Der wesentliche Unterschied zum Korrelationskoeffizienten besteht darin, dass bei der Berechnung des Korrelationskoeffizienten für den Erwartungswert E der Mittelwert der Wortverteilungen genommen wird, bei der α-Abstraktion die Erwartbarkeit vor dem Dokument. Im Falle eines sehr großen Textkorpus geht L gegen ∞ und somit der Erwartungswert gegen null. Somit konvergiert die Riegersche α-Abstraktion in diesem Fall gegen den Cosinus-Koeffizienten (Mehler, 2001, 235) (vgl. S.55).

Für jedes Symbol kann nun der so genannte Korpuspunkt berechnet werden:

$$y_i = (\alpha(x_i,x_j), \ldots \alpha(x_i,x_n)) \tag{3.5}$$

Die Korpuspunkte bilden den **Korpusraum**.

β-Abstraktion Der nächste Schritt ist die paradigmatische (β-)Abstraktion. Der Sinn dieses Schrittes liegt in der Erzeugung eines Raumes, in dem die paradigmatischen Verwendungsregularitäten der Worte widerspiegelt werden. Für die paradigmatische Abstraktion existieren verschiedene Ansätze, die ursprüngliche Definition von Rieger (Rieger, 1981) lautet:

$$\delta(y_i,y_j) = \sqrt{\sum_{n=1}^{W}(\alpha_{i,n} - \alpha_{n,i})^2}; \delta \in [0; 2\sqrt{W}] \tag{3.6}$$

Somit handelt es sich bei der von Rieger vorgeschlagenen β-Abstraktion um eine Euklidische Metrik. Diese Metrik berechnet die syntagmatischen Regularitäten auf dem Korpusraum. Projiziert man den resultierenden Raum durch Division durch $2\sqrt{n}$ in einen normalisierten Raum, erhält man den **Bedeutungsraum**, der die paradigmatischen Verwendungsregularitäten widerspiegelt. Durch die erneute Anwendung einer Euklidischen Metrik auf dem Bedeutungsraum können die paradigmatischen Verwendungsregularitäten gemessen werden.

In der vorliegenden Arbeit werden die Euklidischen Metriken durch das jeweilige gewählte Ähnlichkeitsmaß (üblicherweise den Korrelationskoeffizienten) ersetzt, um so größere Vergleichbarkeit zu schaffen. Auf diese Weise kommt in dieser Arbeit nur ein stark vom ursprünglichen Verfahren abweichender semantischer Raum nach Rieger zum Einsatz.

3.4.3 LSA

In ihrem Aufsatz „A Solution to Plato's Problem" (Landauer und Dumais, 1997) stellten die Autoren einen Lösungsansatz dar, der die Frage beantworten sollte, wie Menschen trotz der geringen Mengen an Informationen, die sie über ihre Sinnesorgane aufnehmen, große Mengen an Wissen sammeln können. Eine Lösung dieser Frage wurde anhand des Problems des Spracherwerbs dargestellt. Obwohl heute wohl niemand mehr der Überzeugung ist, dass das vorgestellte Verfahren dem menschlichen Lernen entspricht, stellt es doch ein Standardverfahren im Bereich des *information retrieval* dar. Ebenso wie bei dem semantischen Raum nach Rieger werden durch die LSA semantische Ähnlichkeitsbeziehungen von Worten betrachtet. Dies geschieht über die Annahme einer „versteckten" – latenten – Struktur, die sich hinter den Wortverwendungsverteilungen verbirgt. Grundsätzlich hat die LSA den Anspruch, die semantischen Beziehungen innerhalb dieses latenten Raumes darzustellen. Ich möchte die Idee des Verfahrens anhand der oben beschriebenen Term-Dokument-Matritzen verdeutlichen, zumal auch die Experimente, die in (Landauer und Dumais, 1997) dargestellt sind, mit solchen Matritzen arbeiten.

Die Idee der LSA besteht darin, unübersichtliche Datenmengen, die aufgrund ihrer Struktur einen Datenraum aufspannen, in einen kleiner dimensionierten Raum zu projizieren[14]. Betrachtet man beispielsweise eine Term-Dokument-Matrix A, so entspricht die Dimensionalität des durch diese Matrix aufgespannten Vektorraums der Anzahl von linear unabhängigen Dokumenten, also der Anzahl der linear unabhängigen Dokumentvektoren in einem Raum, der durch die transponierte Matrix A^T aufgespannt wird. Üblicherweise ist die Dimensionalität r eines Raumes A gleich der oder geringfügig kleiner als die Anzahl der Dokumente D.

Die Reduktion der Dimensionalität wird mit Hilfe einer *singular value decomposition* (SVD) durchgeführt (vgl. (Leopold, 2005), (Landauer und Dumais, 1997) und (Deerwester u. a., 1990)). Dabei handelt es sich um ein Verfahren aus der Linearen Algebra, das eine beliebige Matrix in drei Komponenten zerlegt, für die folgendes gilt:

$$A = UsV \tag{3.7}$$

[14] An dieser Stelle sei angemerkt, dass die LSA in vielerlei Hinsicht mit der PCA (principal components analysis) verwandt ist.

U ist eine $W \times r$ -Matrix mit orthonormalen Spaltenvektoren, V eine $r \times r$ -Matrix mit orthonormalen Zeilenvektoren. Bei s handelt es sich um eine $r \times r$ -Diagonalmatrix, deren Elemente nach absteigender Größe sortiert sind. Sie entsprechen den Singulärwerten von A. Der Theorie der LSA zufolge entspricht jeder Singulärwert einem semantischen Konzept, welches nicht offensichtlich (also *latent*) die Entstehung der Datenmenge beeinflusst hat. Die Größe des Singulärwertes entspricht der Stärke, mit der dieses Konzept Einfluss auf die Entstehung des Korpus genommen hat. Durch die Löschung der kleinsten Singulärwerte können die potentiell störenden Einflüsse der am wenigsten vertretenen Konzepte eliminiert werden, so dass es einfacher ist, die verbleibenden, bedeutenderen Konzepte zu erkennen. Bildlich ausgedrückt werden auf diese Weise Streuvektoren, die möglicherweise auf Messfehler etc. zurückzuführen sind, mit Gruppen von anderen Vektoren gebündelt.

Der von einer LSA aufgespannte Vektorraum entspricht dem dimensionsreduzierten Raum A_K, der durch Multiplikation der reduzierten Matritzen U_K, s_K und V_K entsteht.

$$A_K = U_K s_K V_K \tag{3.8}$$

Die reduzierten Matritzen ergeben sich durch die Löschung der $r - K$ kleinsten Diagonalwerte, bzw. Spalten und Reihen.

Der gewonnene Vektorraum kann ansonsten wie ein einfacher Vektorraum behandelt werden. So können z.B. vor der Durchführung der SVD auf die Term-Dokument-Matrix Entropie/*idf*-Gewichte[15] angewandt (Landauer und Dumais, 1997), bzw. nach der SVD Distanzen zwischen den Wortvektoren des neuen Raumes berechnet werden.

Parametrisierung Eine wichtige Frage, die vor jeder LSA zu klären ist, ist die Frage nach der Stärke der Reduktion, also der konkreten Anzahl von Dimensionen, auf die der Ursprungsraum reduziert werden soll. Auf diese Frage gibt es keine grundsätzliche Antwort, da sie Wissen über den Anteil an Streuvektoren innerhalb des zu analysierenden Korpus und die Größe des Korpus voraussetzt. Oftmals werden in Abhängigkeit von der Größe der Korpora Standardwerte im Bereich von 100-400 Dimensionen gewählt. Trotzdem gibt es Versuche, die optimale Dimensionalität möglichst genau zu approximieren (Kim u. a., 2003).

Für Analysen im Rahmen dieser Arbeit sind Daumenregeln wie die beschriebene Reduktion auf 100-400 Dimensionen bei mittelgroßen Korpora nicht anwendbar, da schon die ursprünglichen Räume keine derart hohe Dimensionalität aufweisen. Eine Grundidee des Ansatzes besteht aber gerade darin, ein Verfahren zu entwickeln, welches schon bei minimalen Trainingsdatenmengen funktioniert. Prinzipiell ist dieses Ziel auch mit einer LSA erreichbar, wie das in (Landauer und Dumais, 1997, S.238ff) beschriebene Spielzeugbeispiel zeigt, in dessen Kontext zwölf Begriffe aus neun Sätzen thematisch sortiert werden konnten. Allerdings wurde im Rahmen dieses Beispiels die Anzahl der Dimensionen willkürlich bestimmt.

Um relativ zur Größe des Trainingskorpus eine Schätzung der optimalen Dimensionalität zu gewinnen, wurde eine Funktion (Folge) in Abhängigkeit von der Anzahl der Trainingsdokumente gewählt, die die bekannten Eckpunkte – stets kleiner als die Anzahl der Dokumente so wie ca. 400 Dimensionen bei großen Korpora – approximiert.

[15] vgl. 3.31 auf Seite 64

Diese Funktion lautet wie folgt:

$$f(x) = \begin{cases} 2 & : \quad x = 2 \\ \lfloor \frac{\sqrt{x}}{1.5} + 0.5 \rfloor & : \quad x > 2 \end{cases} \tag{3.9}$$

Eine Reduktion der Dimensionalität tritt folglich erst bei einer Dokumentanzahl größer als zwei ein, die aber den Standardfall darstellt. Diese Formel wurde in allen in dieser Arbeit durchgeführten Anwendungen der LSA zur Abschätzung der Dimensionsanzahl verwendet.

Kritik An der LSA wurde in verschiedener Hinsicht Kritik geübt; eine davon wurde in (Ando, 2000) geäußert: Die Reduktion der Dimensionalität resultiert darin, dass sowohl Wortvektoren als auch Dokumentvektoren (in der transponierten Matrix) in lineare Abhängigkeit gebracht werden. Auf diese Weise werden sowohl latente Einflüsse auf der Dokumentebene als auch auf der Wortebene gleichermaßen reduziert. Ando schlägt ein Verfahren vor, bei dem dieser Effekt weitgehend auf die Dokumentvektoren begrenzt wird. Anhand verschiedener korpusbasierter Tests konnte er nachweisen, dass das Verfahren durchschnittlich bessere Ergebnisse liefert, als eine normale LSA.

In dieser Arbeit wurde auf die Anwendung von Andos Verfahren aus zweierlei Gründen verzichtet: Zum einen hätte die Umsetzung aufgrund der unzureichenden Beschreibung des Verfahrens, welches im proprietären Kontext entwickelt wurde, zu viel Zeit gekostet. Zum anderen ist unklar, ob Andos Aussage, dass Streuvektoren auf Wortebene wenig Relevanz für *retrieval*-Aufgaben haben, für das dieser Arbeit zugrundeliegende Szenario ebenso gilt[16].

Von der Perspektive der Fuzzy Semantics aus kann ein weiterer Aspekt der LSA kritisiert werden: Die syntagmatischen Beziehungen der Wortvektoren werden nicht eigenständig nach Anwendung einer α-Funktion abgebildet, da die α-Funktion und die β-Funktion gewissermaßen in einem Schritt angewandt werden (Mehler, 2001, S.82). Auf diese Weise verwischen die Grenzen des zweistufigen Prozesses der Analyse der syntagmatischen und paradigmatischen Regularitäten. Dies stellt jedoch kein Hindernis für die Anwendbarkeit der LSA im Kontext der automatischen Indizierung oder Themenerkennung dar (ebd.), so dass an dieser Stelle auf diese Kritik nicht weiter eingegangen wird.

Weitere Kritik an der LSA ergibt sich aus der mangelnden statistischen Fundierung des zugrundeliegenden Verfahrens so wie der impliziten Annahme, dass das von der LSA zu bereinigende Rauschen additiv-Gaus'scher Natur sei, während ein multinomiales Modell eher linguistischen Tatsachen entsprechen würde (Leopold, 2005, S.13). Diese Kritiken führten zu der Entwicklung der wohlfundierten *probabilistic latent semantic analysis*, die ich im folgenden Abschnitt beschreiben möchte.

3.4.4 Probabilistic Latent Semantic Analysis (PLSA)

Die PLSA wurde ursprünglich von Hofmann (Hofmann, 1999) entwickelt. Das Verfahren projiziert Dokumentvektoren in einen semantischen Raum, der durch k latente

[16] Tatsächlich scheint der starke Einfluss der C-Werte auf die Themenerkennungsqualität, der in den *offline*-Experimenten (siehe Kapitel 5) ermittelt werden konnte, nahezulegen, dass auch oder gerade streuende Wortvektoren einen negativen Einfluss haben.

probabilistische Variablen aufgespannt wird. Wie Hofmann zeigen konnte, liefert die PLSA in vielen Fällen bessere Ergebnisse als die LSA.

Wie bei der LSA wird eine Reihe von latenten Faktoren angenommen, welche die Entstehung des Korpus beeinflusst haben, ohne dabei selbst direkt beobachtbar zu sein. Diese Faktoren werden durch die erwähnten latenten Variablen z_1 bis z_K repräsentiert.

Die Wahrscheinlichkeit, dass ein Wort und ein Dokument gemeinsam produziert werden, entspricht:

$$P(d, w) = P(d)P(w \mid d) \tag{3.10}$$

Die Wahrscheinlichkeit, dass ein Wort w in einem Dokument d produziert wird, ist unter der Annahme von K latenten Variablen:

$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)p(z_k \mid d_j) \tag{3.11}$$

Der semantische Raum, der durch eine PLSA aufgespannt wird, wird durch die Wahrscheinlichkeiten $p(z_k \mid d_j), k = 1, \ldots, K$ definiert, wobei es sich allerdings um eine Projektion der Dokumentvektoren (und nicht der Wortvektoren) handelt.

Die Wahrscheinlichkeiten $p(z \mid d)$ und $p(w \mid z)$ werden anhand einer Term-Dokument-Matrix, die zum Training verwendet wird, ermittelt. Da die Einflüsse der latenten Variablen nicht direkt beobachtbar sind, wird dazu zur Approximation ein iterativer Lernalgorithmus, das *expectation maximization*-Verfahren (Dempster u. a., 1977) benutzt.

Durch die Anwendung des Bayes'schen Gesetzes auf die aufgrund der angenommenen, multinomialen Verteilung berechneten *likelihood*-Funktion wird die *expectation*-Funktion des Algorithmus gewonnen[17]:

$$p(z_k \mid w_i, d_j) = \frac{p(w_i \mid z_k)p(z_k \mid d_j)}{\sum_{k=1}^{K} p(w_i \mid z_k)p(z_k \mid d_j)} \tag{3.12}$$

In diese Gleichung werden zufallsbasierte Werte für $p(w_i \mid z_k)$ und $p(z_k \mid d_j)$ eingesetzt, die allerdings stets aufsummiert jeweils 1 ergeben müssen, so wie größer oder gleich 0 sind.

Die berechneten Werte für $p(z_k \mid w_i, d_j)$ werden in die beiden Formeln eingesetzt, um neue, besser approximierte Werte für $p(w_i \mid z_k)$ und $p(z_k \mid d_j)$ zu erhalten (*maximization*):

$$p(w_i \mid z_k) = \frac{\sum_{j=1}^{N} f(w_i, d_j)p(z_k \mid w_i, d_j)}{\sum_{k=1}^{K}\sum_{j=1}^{D} f(w_i, d_j)(z_k \mid w_i, d_j)} \tag{3.13}$$

$$p(z_k \mid d_j) = \frac{\sum_{i=1}^{W} f(w_i, d_j)p(z_k \mid w_i, d_j)}{f(d_j)} \tag{3.14}$$

Die neu berechneten Werte werden wiederum in Gleichung 3.12 eingesetzt und der Prozess wiederholt, bis sich keine Veränderungen mehr ergeben.

Wie geschildert lassen sich mit der PLSA oft bessere Ergebnisse erreichen, als mit der LSA. Eine Kritik an dem Verfahren besteht allerdings in der Verwendung des

[17] Für eine detaillierte Herleitung vgl. (Leopold, 2005, S.10ff), eine detaillierte Beschreibung des Gesamtprozesses findet sich unter (Hofmann, 2001).

EM-Algorithmus, der als Suchalgorithmus nur lokale Maxima zu finden in der Lage ist, so dass ggf. suboptimale Werte für die bedingten Wahrscheinlichkeiten ermittelt werden.

Auch wenn die PLSA für zukünftige Forschungen im Bereich der Themenerkennung für situierte Mensch-Roboter-Kommunikation potentiell sehr nützlich sein wird, wurde im Rahmen dieser Arbeit auf eine Anwendung dieses Verfahrens verzichtet, da die Struktur des resultierenden semantischen Raumes von der der ersten drei Verfahren stark abweicht, und so direkte Vergleiche potentiell problematisch sind. Ein weiteres Problem besteht darin, dass der EM-Algorithmus ausgehend von den zufällig gewählten Initialisierungswerten nur lokale Maxima findet, also das Verfahren – insbesondere angesichts der kleinen Testmengen – nicht deterministisch evaluierbar wäre. Ggf. liesse sich aber dieses Problem durch eine nicht-zufallsbasierte Initialisierung beheben. Zuletzt müsste das Verfahren der PLSA so modifiziert werden, dass eine Projektion der Wortvektoren vorgenommen werden würde – wie oben geschildert ist das Verfahren für die Projektion von Dokumentvektoren entwickelt worden. Zukünftige Forschungen sollten aber versuchen, die PLSA für das zugrundeliegende Forschungsthema nutzbar zu machen.

3.4.5 Weitere semantische Räume

Neben den beschriebenen vektorraumbasierten Verfahren existieren verschiedene weitere Verfahren, die die semantischen Beziehungen von Worten verteilungsbasiert zu erfassen versuchen. Stichpunktartig seien hier Selbstorganisierende Karten (SOM – Self Organizing Maps) erwähnt. Neben dem ursprünglichen Modell (Kohonen, 1981) finden sich viele Erweiterungen, darunter die Hyperbolischen Selbstorganisierenden Karten (Ontrup und Ritter, 2005) (HSOM), die mit Erfolg zur dynamischen Einordnung (Klassifikation) hochdimensionaler (Wort-)Vektoren eingesetzt wurden. Selbstorganisierende Karten kamen im Kontext dieser Arbeit allerdings nicht zum Einsatz, da die Anzahl der Klassen für dieses Verfahren vorgegeben werden muss, was der Anforderung nach Dynamizität widerspricht.

Edda Leopold diskutiert in (Leopold, 2005) einen weiteren Typ von semantischen Räumen, der durch den Einsatz von Klassifikatoren wie dem Bayes-Klassifikator oder Support-Vektor-Maschinen (SVM) (Schölkopf und Smola, 2002) gewonnen wird. In diesen Räumen wird anhand einer Trainingsmenge ein klassenspezifischer[18] Klassifikator trainiert, der dann (unter Berücksichtigung einer gewissen Fehlerrate) entscheiden kann, ob Vektoren zu dieser gehören oder nicht.

Der Nachteil klassifikatorbasierter semantischer Räume entspricht dem der Selbstorganisierenden Karten und geht sogar darüber hinaus: nicht nur die Anzahl der Kategorien (Themen) muss spezifiziert werden, sondern für jeden Klassifikator muss eine eigene Trainingsmenge thematisch relevanter und nicht-relevanter Dokumente erstellt werden. Somit verletzten diese semantischen Räume erneut das Dynamizitätskriterium.

Allerdings will ich an dieser Stelle nicht verschweigen, dass prinzipiell auch dynamische Themenerkennung mit klassifikator-basierten Ansätzen möglich ist. Dazu muss nur in einem ersten Schritt dynamisch eine Trainingsmenge für jeden Klassifikator gebildet werden, anhand derer dieser dann trainiert wird. Natürlich ist fraglich, wie

[18] In der Themenerkennung kann z.B. ein Thema eine Klasse darstellen.

weit die einzelnen Klassifikatoren die dadurch möglicherweise entstehenden Fehler im Trainingsmaterial kompensieren können. Tatsächlich stellt das in dieser Arbeit gewählte Verfahren einen solchen Ansatz dar. Die vorliegende Diskussion behandelt jedoch die Frage, wie weit semantische Räume zum *Training* eines Klassifikators verwendet werden können, und klassifikator-basierte semantische Räume lassen sich nicht ohne Verletzung des Dynamizitätskriteriums für diese Aufgabe einsetzen.

Wie aus dem Gesagten ersichtlich ist, beschränkt sich diese Arbeit auf die Anwendung der basalen, vektorraumbasierten Verfahren. Die Entscheidung für diese Tatsache wurde aus zwei Gründen getroffen: Zum einen liegt der Schwerpunkt dieser Arbeit weniger in der exzessiven Optimierung des Themenerkennungsprozesses als vielmehr in der Untersuchung, welche Randbedingungen für eine Themenerkennung auf einem situiert kommunizierenden Robotersystem herrschen und wie diese utilitarisiert werden können. Zum anderen sind die gewählten Verfahren hinreichend ähnlich, dass eine hohe Modularität des zu entwickelnden Systems möglich wurde. Auf diese Weise können gezielt Änderungen z.B. an dem Distanzmaß vorgenommen werden und die Effekte im Kontext aller semantischen Räume gleichermaßen betrachtet werden.

3.4.6 Ähnlichkeitsmaße

Im vorigen Abschnitt 3.4 wurden verschiedene Modelle vorgestellt, die die Grundlage für ein thematisches Clustern und damit für die Berechnung von semantischen bzw. thematischen Distanzen zwischen Wortvektoren, die für die Aufgabe der *cluster detection* notwendig ist, darstellen können. Ich möchte jetzt auf die Frage nach der Berechnung dieser Distanzen eingehen.

Unter der Annahme eines Euklidischen Vektorraums (also bei den vorgestellten Vektorraummodellen) können zur Berechnung des Abstands[19] zweier Vektoren x und y verschiedene Maße herangezogen werden. Relevant für die Berechnung diverser Distanzmaße ist das **Skalarprodukt** zweier Vektoren, das ich kurz vorstellen möchte:

Definition 3.8 (Skalarprodukt) *Die Multiplikation „$\cdot$“ von zwei Vektoren sei das Skalarprodukt derselben. Das Skalarprodukt zweier Vektoren x und y mit der Dimensionalität n berechnet sich wie folgt (Duda u. a., 2001, S.606):*

$$x \cdot y = x^T y = y^T x = \sum_{i=1}^{n} x_i y_i \tag{3.15}$$

Die Quadrierung x^2 entspricht dem Skalarprodukts eines Vektors mit sich selbst: $x \cdot x$

Die folgende Auflistung stellt einige übliche Distanzmaße vor, wobei allerdings kein Anspruch auf Vollständigkeit erhoben wird:

- Euklid: Die **Euklidische Distanz** berechnet den Abstand zwischen den durch die Vektoren definierten Punkten in einem Vektorraum. Sie entspricht der Euklidischen Norm der Differenz der Vektoren (Bronstein u. a., 2001, S.624):

$$D_{euk}(x, y) = ||x - y|| \tag{3.16}$$

[19] Prinzipiell existiert die Unterscheidung zwischen Distanzmaßen und Ähnlichkeitsfunktionen. Die folgenden Verfahren stellen fast ausschließlich letztere dar. Der Einfachheit halber wird im Folgenden auf diese Differenzierung verzichtet.

Die **Euklidische Norm** (Länge) eines Vektors x ist wie folgt definiert (Bronstein u. a., 2001, S.328):

$$||x|| = \sqrt{x \cdot x} = \sqrt{\sum_{i=1}^{n} {x_i}^2} \tag{3.17}$$

Weswegen gilt:

$$D_{euk}(x, y) = ||x - y|| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.18}$$

- Cosinus: Das **Cosinus-Distanzmaß** (Bronstein u. a., 2001, S.328) ist im Bereich vieler texttechnologischer Untersuchungen ein Standardmaß. Es entspricht dem Cosinus des Winkels der zu vergleichenden Vektoren im Ursprung.

$$D_{cos}(x, y) = \cos \alpha(x, y) = \frac{x \cdot y}{||x||\ ||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3.19}$$

- Korrelationskoeffizient: Dieses Maß dient der Berechnung der statistischen Korrelation zweier Zufallsvariablen. Der so genannte **empirische Korrelationskoeffizient** kann für den Vergleich zweier Messreihen, in diesem Fall von Vektoren, eingesetzt werden (Bronstein u. a., 2001, S.801):

$$D_{kor}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{3.20}$$

  Die Schätzwerte errechnen sich dabei wie folgt:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{3.21}$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \tag{3.22}$$

- Jaccard: Die **Jaccard-Distanz** entspricht der Größe der Schnittmenge zweier Mengen A und B dividiert durch die Größe der Vereinigung derselben ($J(A, B) = |A \cap B| / |A \cup B|$). Binäre Vektoren[20] können wie folgt verglichen werden (Lohninger, 2005):

$$Jac_{bin} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{3.23}$$

  wobei

  - M_{11} die Anzahl der Fälle ist, in denen sowohl x als auch y an einer Stelle den Wert 1 besitzen,

[20] Also solche, die nur 1 oder 0 als Wert an einer Stelle haben können.

  - M_{01} die Anzahl der Fälle ist, in denen x an einer Stelle den Wert 0 und y den Wert 1 besitzt, und
  - M_{10} die Anzahl der Fälle ist, in denen x an einer Stelle den Wert 1 und y den Wert 0 besitzt.

- Tanimoto: Eine Erweiterung auf nicht-binäre Vektoren anhand des Cosinus ist die so genannte **Tanimoto-Distanz** (Duda u. a., 2001, S.541):

$$D_{tan}(x, y) = \frac{x \cdot y}{\|x\|^2 \|y\|^2 - x \cdot y} \tag{3.24}$$

Für die vorliegende Arbeit wurden der Korrelationskoeffizient, die Euklidische Distanz so wie die Cosinus-Distanz implementiert, weiterhin macht der modulare Aufbau der entwickelten Systeme die Integration weiterer möglich. Die Wahl des Ähnlichkeitsmaßes sollte jedoch eher geringen Einfluss auf die möglichen Ergebnisse haben. Grundlage aller Versuche bildete – letztendlich willkürlich gewählt – der Korrelationskoeffizient. Die Hoffnung bestand, dass ein Clusterprozess bei Korrelationswerten von 0 – also unkorreliert – und geringer zwischen Vektoren trennen und so einen Anhaltspunkt für einen dynamischen Clusterprozess haben würde. Außerdem ist der Korrelationskoeffizient schon normalisiert, er kann nur Werte zwischen -1 und 1 annehmen.

3.5 Clusterverfahren

In der modernen Musterklassifikationsforschung existiert eine große Zahl von Verfahren zur Clusterung von Vektorräumen. Aufgrund des modularen Aufbaus, der dem im Rahmen dieser Arbeit erstellten Themenerkennungssystem zugrunde liegt, konnten verschiedene Clusterverfahren implementiert und getestet werden, darunter MCL (van Dongen, 2000) – eigentlich ein Graph-Clusteralgorithmus, dafür aber sehr schnell – *Support Vector Clustering* (SVC) (Ben-Hur u. a., 2001), ein multivariater LBG und verschiedene hierarchisch-agglomerative Clusteralgorithmen.

Sehr schnell stellte sich heraus, dass die hierarchisch-agglomerativen Verfahren – die neben KMeans-basierten Ansätzen wie LBG in der Texttechnologie Standardverfahren sind – die besten Ergebnisse lieferten, so dass in allen Experimenten ausschließlich mit diesen Algorithmen gearbeitet wurde. Die „exotischen“ Algorithmen MCL und SVC stellten sich als zu langsam (SVC) heraus oder lieferten zu schlechte Ergebnisse (MCL), was aber aufgrund des untypischen Anwendungsgebietes erklärbar ist. Mit dem multivariaten LBG wurde ein Verfahren ausprobiert, welches mit der Schätzung von Normalverteilungen in dem Datenraum einen Gegensatz zu den „hart“ clusternden alternativen Verfahren darstellt. Auf diese Weise wird nicht jeder Vektor fest einem Cluster zugewiesen, sondern für jeden Vektor kann die Wahrscheinlichkeit der Zugehörigkeit zu einem Cluster bestimmt werden. Allerdings benötigt die stabile Schätzung von Normalverteilungen eine Datengrundlage einer bestimmten Größe, weswegen das Verfahren zu keinen guten Ergebnissen führte.

Die letztendliche Entscheidung für einen hierarchisch-agglomerativen Algorithmus wurde weiterhin durch folgende Gründe gestützt:

- Hierarchisch-agglomeratives Clustern erlaubt sowohl die Vorgabe einer festen Clusteranzahl als auch die Begrenzung des Clusterungsprozesses durch Spezifikation eines Schwellwertes oder eine Kombination beider Verfahren. Auf diese Weise konnten sowohl Versuche mit einer festen Clusteranzahl als auch Versuche mit echt dynamischer Clusterung durchgeführt werden.

- Hierarchisch-agglomeratives Clustern liefert auch im Vergleich zu KMeans-Varianten wie Bi-Section-KMeans gute Ergebnisse (Cimiano u. a., 2004).

- Der Clustersvorgang ist im Detail nachvollziehbar und im Gegensatz zu anderen Verfahren deterministisch; die Ergebnisse sind also nicht von einer zufälligen Wahl von Initialisierungsparametern abhängig.

- Hierarchisch-agglomeratives Clustern ist mit einer Komplexität von $O(n^2\ log\ n)$[21] (complete linkage) zwar nicht sehr schnell, hat sich aber in den unten beschriebenen Experimenten als ausreichend schnell erweisen, so dass kein Zwang zur Verwendung schnellerer Algorithmen bestand.

Ich möchte im Folgenden das verwendete Clusterverfahren skizzieren.

Hierarchisch-agglomeratives Clustern (Cimiano u. a., 2004) unterteilt die von ihm betrachteten Clusterverfahren, die auf Distanzvergleich von Vektoren basieren, in zwei Kategorien: **agglomerativ** (bottom-up) und **divisiv** (top-down). Divisive Verfahren nehmen die zu clusternde Menge aller Vektoren und unterteilen sie anhand bestimmter Kriterien, bis ein – zu definierender – Endzustand erreicht ist. Agglomerative Verfahren fügen Vektoren und Vektormengen sukzessive zusammen, bis ebenfalls ein Endzustand eintritt. Die Kriterien, anhand derer Vektoren(mengen) zusammengefügt werden, können stark variieren und sind oftmals aufgabenabhängig. Im Rahmen der distanzbasierten Clusterung von Vektorräumen existieren jedoch drei Basisverfahren (Rieger, 1989, S.216ff): *Single linkage*, *complete linkage* und *average linkage*.

Sei V ein Vektorraum bestehend aus m Vektoren $v_1 \dots v_m$. C sei die Menge der Vektoren. Zwischen den Vektoren sei eine Distanzbeziehung $\delta(v_a, v_b) \geq 0$ definiert. $\prod^v$ sei eine Zerlegung der Menge der Vektoren in Teilmengen durch den Clusteralgorithmus, wobei v den Schritt des Clusteralgorithmus bezeichnet, durch den $\prod^v$ gebildet wurde.

Es gilt: $\prod^0 = \{\{v_1\}, \dots, \{v_m\}\}$ und $\prod^v = \{\{v_1, \dots, v_m\}\}$, für jeden Schritt v werden exakt zwei Teilmengen aus der Zerlegung $\prod^{v-1}$ miteinander vereinigt. Somit nimmt v nacheinander alle ganzzahligen Werte von 0 (vor Anwendung des Algorithmus) bis $m - 1$ an.

Eine vollständige Clusterung eines Vektorraums kann als Baum dargestellt werden, der Aufschluss über die Beziehungen der Vektoren untereinander gibt. Aus diesem Grund wird das Verfahren als hierarchisch bezeichnet.

Es stellt sich die Frage, welche beiden Teilmengen A_p und A_q aus einem $\prod^{v-1}$ miteinander vereinigt werden, um eine Vereinigungsmenge A_k zu erhalten, die A_p und A_q in $\prod^v$ substituiert. Die Antwort auf diese Frage ist durch die Wahl des Verfahrens bedingt. Wie geschildert existieren drei übliche Vorgehensweisen, diese beiden Mengen zu bestimmen.

[21] ebd.

Single linkage Für das *single linkage*-Verfahren werden die Distanzen aller Vektoren zweier verschiedener Mengen einer Zerlegung betrachtet. Die *single linkage*-Distanz zweier Mengen entspricht der kleinsten dieser Distanzen:

$$D^s_{A_i,A_j} =_{def} \min_{y \in A_i, y' \in A_j} \delta(y, y'); y, y' \in C \tag{3.25}$$

Betrachtet man jetzt die Menge P aller $\{(A_i,A_j)\}$, so wird genau das Mengenpaar (A_p,A_q) miteinander vereinigt, welches die kleinste *single linkage*-Distanz besitzt:

$$(A_p, A_q) =_{def} \min_{A_i \neq A_j} \{D^s_{A_i,A_j}\}; A_i, A_j \in \prod^{v-1} \tag{3.26}$$

Die Vereinigung der beiden Mengen führt wie beschrieben zur nächsten Zerlegung $\prod^v$.

Complete linkage Betrachtet man die Distanzen aller Vektoren zweier Mengen zueinander, so entspricht die *complete linkage*-Distanz dem größten dieser Abstände:

$$D^c_{A_i,A_j} =_{def} \max_{y \in A_i, y' \in A_j} \delta(y, y'); y, y' \in C \tag{3.27}$$

Während eines Clusterschrittes werden wiederum die beiden Mengen A_p und A_q miteinander vereinigt, die die geringste Distanz zueinander besitzen:

$$(A_p, A_q) =_{def} \min_{A_i \neq A_j} \{D^c_{A_i,A_j}\}; A_i, A_j \in \prod^{v-1} \tag{3.28}$$

Durch die Vereinigung der beiden Mengen entsteht wie gehabt $\prod^v$.

Average linkage Die *average linkage*-Distanz zweier Mengen von Vektoren entspricht dem durchschnittlichen Abstand dieser Vektoren zueinander:

$$D^a_{A_i,A_j} =_{def} \frac{1}{m_i m_j} \sum_{y \in A_i}^{m_i} \sum_{y' \in A_j}^{m_j} \delta(y, y'); y, y' \in C \tag{3.29}$$

wobei m_i bzw. m_j der Anzahl der Vektoren in A_i respektive A_j entspricht.

Die Durchführung des eigentlichen Clusterns geschieht genau wie bei den anderen Verfahren durch die Vereinigung von A_p und A_q, wodurch $\prod^v$ entsteht.

$$(A_p, A_q) =_{def} \min_{A_i \neq A_j} \{D^a_{A_i,A_j}\}; A_i, A_j \in \prod^{v-1} \tag{3.30}$$

Im Rahmen dieser Arbeit wurde stets das *average linkage*-Verfahren verwendet, da es potentiell die besten Ergebnisse erzeugt.

3.6 Klassifikation

Das Resultat des Clusterprozesses ist eine Gruppierung aller Wortvektoren in dem gewählten semantischen Raum, wobei angenommen wird, dass jede Gruppe ein Thema widerspiegelt. Das Ziel des nachfolgenden Schrittes – der Klassifikation – ist die Zuordnung von neuen Dokumentvektoren zu den Themen. Zur Erinnerung: Die zu klassifizierenden Dokumente und die Trainingsdokumente unterscheiden sich im Rahmen dieser Arbeit in fast allen Fällen durch ihre Segmentierung voneinander. Im Prototypensystem sind die kleinsten behandelten Dokumente die Benutzeräußerungen (*utterances*), also Segmente von kontinuierlich gesprochener Sprache. Obwohl es prinzipiell möglich wäre, diese zum Aufbau eines semantischen Raumes zu verwenden, wäre es dennoch problematisch, da sie nur wenige Worte und damit wenig Kookkurrenzinformation enthalten. Das Training geschieht also mit Segmenten, die anhand eines Segmentierungsprozesses[22] aus konsekutiven Äußerungen gewonnen werden.

Um möglichst schnell auf mögliche Themenwechsel einzugehen[23] werden jedoch die Äußerungen thematisch klassifiziert.

Der direkte Weg zur Klassifikation hätte – wie in Abschnitt 3.3 auf Seite 40 geschildert – in einer Gruppierung der Dokumentvektoren (also Segmente) nach Themen resultiert. Die Gruppen von Dokumentvektoren hätten in diesem Fall die Trainingsmenge für einzelne Klassifikatoren dargestellt, wie z.B. für Bayes-Klassifikatoren oder Support-Vektor-Maschinen.

Für die Klassifikation neuer Dokumente (also Äußerungen) anhand der Gruppen von Wortvektoren wurde ein sehr einfacher Klassifikator gewählt. Dieser Klassifikationsprozess lässt sich zweifellos noch verbessern, diese Aufgabe stand aber nicht im Fokus der vorliegenden Arbeit.

Für jeden (Wort- bzw. Symbol-)Vektor v_n aus dem Vektorraum und für jedes Thema T wurde die durchschnittliche Distanz des Vektors zu allen Vektoren des Themas berechnet. Das Resultat ist eine Liste von Distanzen von jedem Vektor zu jedem Thema.

Für jedes bekannte Thema und alle bekannten Worte aus einem neuen, zu klassifizierenden Dokument werden die Distanzen aufsummiert. Das Thema mit der geringsten[24] aufsummierten Distanz ist das Ergebnis des Klassifikationsvorgangs. Ein Vorteil dieses Klassifikators besteht darin, dass er im Verwendungskontext ausreichend schnell ist. Der gesamte Klassifikationsprozess ist in $O(themenanzahl \cdot äußerungslänge \cdot log(wortanzahl))$ abgeschlossen, wobei die Themenanzahl üblicherweise klein ist[25]. Auf diese Weise können neu eingehende Äußerungen in Quasi-Echtzeit klassifiziert werden.

Rückweisung Bei der Klassifikation ist es möglich, Äußerungen in eine Rückweisungsklasse einzusortieren, wenn die (gemittelte) Summe der Distanzen einen bestimmten Wert unter- bzw. überschreitet. Gerade beim Korrelationskoeffizienten stellt der Wert 0 eine kritische Grenze dar, so dass z.B. alle Dokumente für die

[22] Eigentlich: Unifikationsprozesses

[23] Ansonsten würde ein Thema erst erkannt werden, wenn es abgeschlossen ist.

[24] Dies entspricht nicht immer dem geringsten numerischen Wert. Im Fall des Korrelationskoeffizienten entspricht die höchste Summe der geringsten Distanz.

[25] Unter der Annahme eines *hashing*-Algorithmus, der Suchvorgänge in logarithmischer Abhängigkeit ermöglicht.

kein Thema gefunden wurde, dessen Distanzsumme größer als 0 ist, als nicht klassifizierbar angesehen werden können. Im Falle der Anwendung der Themenerkennung auf einem Robotersystem würden diese Fälle z.B. Situationen entsprechen, in denen sich der Roboter über das aktuelle Thema unsicher ist und nachfragt, anstatt nur zu schätzen.

History Durch das zeitnahe Klassifizieren von Äußerungen entsteht das Problem, dass der Klassifikationsprozess fehleranfällig wird, wenn einzelne Worte ein falsches Thema anzeigen. Ein Beispiel wäre eine Unterhaltung über eine Sofaecke, die sich neben der Küche befindet. Eine Äußerung könnte aber Elemente der Küche mit einbeziehen („gleich vor dem Küchentisch“), weswegen für diese Äußerung fälschlicherweise das Thema „Küche“ erkannt werden würde.

Um diesen Effekt zu kompensieren, kann Verlaufsinformation in Form einer *history* in den Klassifikationsvorgang mit einfließen. Dazu werden die aufsummierten Distanzen zu allen Themen der letzten Äußerung mit einem Verfallsfaktor multipliziert – im Rahmen dieser Arbeit wurde hierfür stets 0.3 gewählt – und mit in die Summenbildung für die aktuelle Äußerung aufgenommen.

Ein Vorteil dieses Verfahrens ist, dass im Falle eines erkannten Themenwechsels – also des Endes eines Segmentes – die *history* gelöscht werden kann bzw. sie zur Klassifikation der nachfolgenden Äußerung herangezogen wird. Zu beachten ist weiterhin, dass in Fällen, in denen kein themenanzeigendes Symbol in einer Äußerung gefunden werden konnte, automatisch das letzterkannte Thema als weiterhin gültig anerkannt wird. Im Prototypensystem wurde angesichts solcher Äußerungen der Verfallsfaktor auf 1.0 gesetzt, um den Einfluss von Füllphrasen auf die Themenerkennung zu unterbinden.

3.7 Exkurs: Vorverarbeitung

Nachdem ich in den vorigen Abschnitten die wesentlichen Elemente eines auf semantischen Räumen basierenden Verfahrens zur Themenerkennung skizziert habe, möchte ich auf die möglichen Störfaktoren, die ein solches Verfahren negativ beeinflussen können, eingehen und Mittel zu ihrer Reduzierung diskutieren. Dies soll auch in Hinblick auf die Ansatzpunkte geschehen, an denen eine multimodale Realisierung eines solchen Verfahrens Verbesserungen gegenüber einer unimodalen Realisierung erzielen kann. Im anschließenden Unterkapitel werde ich die konkreten Verbesserungen diskutieren, die durch eine multimodale, situierte Realisierung im Rahmen dieser Arbeit erzielt wurden.

Wie in Abschnitt 3.4 geschildert, ist die grundlegende Informationsquelle für das unüberwachte Konstruieren von semantischen Räumen die Information über Kookkurrenz von Zeichen in Kontexten (Kommunikationssegmenten, Zeitungsartikeln, etc.). Es existieren eine Reihe von Störfaktoren, die einen Trackingvorgang und/oder die Qualität eines semantischen Raumes als Trainingsbasis für einen Klassifikationsprozess verschlechtern können. Diese sind im Wesentlichen:

1. semantisch/thematisch irrelevante Wörter

2. Polysemie/Polytextie

3. Synonymie
4. zu seltene Wörter
5. suboptimale Segmentierung

Ich werde diese Faktoren im Einzelnen darlegen und Strategien zur Vermeidung diskutieren. Dabei werde ich der Einfachheit halber in den Beispielen auf die Konstruktion eines semantischen Raumes zur Analyse von thematischen Beziehungen eingehen, obgleich diese Störfaktoren in vielen Fällen relevant für Textklassifikationsvorgänge im Allgemeinen sind.

Semantisch/thematisch irrelevante Wörter Ein Störfaktor sind Wörter, die unabhängig oder weitgehend unabhängig von den semantischen oder thematischen Ähnlichkeitsbeziehungen in der Datengrundlage vorkommen. Ein Beispiel dafür sind Funktionswörter wie z.B. Konjunktionen oder Artikel. Allerdings können auch Nicht-Funktionswörter thematisch irrelevant bzw. schwach relevant sein, insbesondere dann, wenn sie für die Datengrundlage typisch sind. So könnte z.B. die Datenbank der Nachrichtenartikel eines Wirtschaftsmagazins sehr häufig das Wort „*money*“ enthalten, welches z.B. nur wenig Rückschlüsse auf das jeweilige Thema (Ölhandel, Aktienkurse, etc.) zulässt. In einem anderen Kontext – z.B. einer Tageszeitung – würde es jedoch ein starkes Indiz für ein wirtschaftsbezogenes Thema sein.

Der Umstand, dass häufig vorkommende Wörter nur einen irrelevanten Beitrag zur Textklassifikation leisten können, wurde schon früh von Luhn (Luhn, 1958) entdeckt. Dieser fand heraus, dass bei einer Ranglistensortierung der in einem Textkorpus vorkommenden Worte insbesondere Worte mittleren Rangs Aussagen über die Textsorte machen. Somit stellen sowohl häufig als auch selten vorkommende Worte einen Störfaktor bei verteilungsbasierten semantischen Analysen dar. Funktionswörter – als ein Fall von zu häufig vorkommenden Wörtern – werden im Kontext des *information retrieval* und des *text mining* in den meisten Fällen über eine sog. „Stoppliste“ zu ignorierender Wörter ausgeschlossen. Andere schwach relevante Worte werden üblicherweise über ihre Verteilung bestimmt (und entfernt). Ich werde im Abschnitt „Gewichte“ detaillierter auf diesen Punkt eingehen.

Zu seltene Wörter Bei dieser Gruppe von Wörtern ist aufgrund einer zu geringen Datenmenge keine sichere Zuweisung zu einem Thema möglich. Die Positionierung ihres Wortvektors in einem semantischen Raum ist somit sehr wahrscheinlich fehlerbehaftet. Obgleich dieser Umstand für semantische Räume und Textklassifikationsverfahren im Allgemeinen problematisch sein kann, stellt er jedoch im Rahmen der meisten Ansätze zur Textklassifikation kein Problem dar, da das seltene Vorkommen dieser Worte ihren Einfluss auf Klassifikationsprozesse stark begrenzt. Ein Grund dafür liegt in der üblichen Länge von zu klassifizierenden Dokumenten, in denen einzelne Vorkommen von schlecht klassifizierbaren Symbolen angesichts einer Vielzahl gut klassifizierbarer wenig stören. Da im Rahmen dieser Arbeit die zu klassifizierenden Äußerungen jedoch extrem kurz sind, können zu selten vorkommende Wörter einen stark störenden Einfluss auf die Klassifikation der Äußerungen, in denen sie vorkommen, ausüben. Weiterhin stellen sie einen Störfaktor für den Clustervorgang dar.

Aus diesem Grund sollte ein Mechanismus gefunden werden, der Fehler aufgrund dieser Worte ausschließt. Dies geschieht in dieser Arbeit einfach durch die Definition einer Mindestanzahl von Dokumenten (Segmenten), in denen ein Wort vorkommen muss, um in den semantischen Raum aufgenommen zu werden. Diese Schwelle wird im Folgenden mit C bezeichnet.

Polysemie/Polytextie Polyseme – also Zeichen mit mehreren Bedeutungen – können ebenfalls problematisch sein, wenn die Einzelbedeutungen unterschiedlichen Themengebieten zuzuordnen sind. Allgemeiner und von der Semantik der Wörter abstrahierend definiert Köhler in (Köhler, 1986, S.63) den Begriff der **Polytextie**:

Definition 3.9 (Polytextie (Köhler)) *Die POLYTEXTIE einer lexikalischen Einheit ist die Anzahl der Kontexte, in denen sie verwendet wird.*

Die Polytextie ist nach Köhler Resultat des Bedürfnisses, Lexeme kontextunabhängig verwenden zu können, da sich sonst die Verwendung von Sprache unökonomisch gestaltet[26]. Objektreferenzen haben im Allgemeinen eine sehr hohe Polytextie, da sie in unzähligen Situationen ein unterschiedliches Denotatum haben können. Ebenso haben Funktionswörter eine hohe Polytextie, da sie in einer großen Anzahl von Kontexten Verwendung finden. Polytextie und Polysemie (bzw. Polylexie[27]) sind miteinander korreliert[28], so dass eine Reduktion der Zahl der Polyseme innerhalb eines Textes üblicherweise die durchschnittliche Polytextie senkt.

Im Kontext des im Rahmen dieser Arbeit dargestellten Themenerkennungsprozesses erzeugen Begriffe mit hoher Polytextie oft das Problem, dass während eines Clustervorgangs thematisch nicht zusammengehörende Cluster verschmolzen werden, da sie durch den Begriff miteinander verbunden sind. Auf diese Weise sind Funktionswörter doppelt problematisch: Sie tragen nicht nur wenig zur Erkennung der Themas bei (s.o.), sondern weisen auch eine hohe Polytextie auf.

Hohe Polytextie stellt im Rahmen dieser Arbeit aber nur dann ein Problem dar, wenn die lexikalischen Einheiten innerhalb verschiedener *thematischer* Kontexte Verwendung finden. Es kann auch nicht das Ziel sein, die Polytextie aller lexikalischen Einheiten z.B. durch geeignete Substitutionen auf das Minimum – eins – zu reduzieren, da ansonsten das Problem der zu selten vorkommenden Wörter eine Themenerkennung unmöglich machen würde. Ziel sollte also eine gezielte Differenzierung von Wortverwendungen sein, so dass im Idealfall jede lexikalische Einheit nur im Rahmen eines thematischen Kontexts verwendet wird.

Es gibt verschiedene Ansätze, Teilbedeutungen zu erkennen und voneinander zu trennen (also die Polysemie/Polylexie zu reduzieren); einige dieser Ansätze – wie z.B. (Schütze, 1998) – arbeiten sogar auf der Basis semantischer Räume. Im Rahmen dieser Arbeit wird Polytextie speziell bei Objektbezeichnungen multimodal aufgelöst. Auf diesen Vorgang wird in Abschnitt 3.8.2 auf Seite 66 genauer eingegangen.

[26] Z.B. müsste im Extremfall für jedes referenzierte Einzelobjekt ein Eigenname entwickelt werden.

[27] Die **Polylexie** ist nach (Köhler, 1986, S.63): „Die Anzahl der verschiedenen Bedeutungen, die eine lexikalische Einheit zu einem gegebenen Zeitpunkt trägt (...)". Im Gegensatz zum Begriff der Polysemie wird bei Köhler nicht zwischen semantischen und grammatischen Bedeutungen unterschieden.

[28] In Bezug auf die Polylexie, vgl. (Köhler, 1986). Bezüglich der Polysemie unbewiesen, es wird an dieser Stelle als plausibel angenommen.

Synonymie Der Begriff der Synonymie kennzeichnet im Allgemeinen Wörter bzw. Zeichen, die trotz unterschiedlicher lexikalischer Repräsentation dieselbe Bedeutung haben. Da einem semantischen Raum, der aus lexikalischen Repräsentationen von Zeichen aufgebaut wird, die Information über Synonymie nicht zur Verfügung steht, werden für die einzelnen Synonyme einzelne Vektoren gebildet. Im Fall großer Datenmengen stellen insbesondere strikte Synonyme (Orange - Apfelsine) für den Aufbau semantischer Räume ein eher geringes Problem dar, da die jeweiligen Vektoren in unmittelbarer Nachbarschaft verortet werden. Trotzdem bedeutet Synonymie in jedem Fall den Verlust an Kookkurrenzinformation, so dass in Textklassifikationssystemen z.B. durch Lemmatisierung, Stammformenbildung oder Thesauri Synonyme vereinigt werden. Im Rahmen dieser Arbeit dient die Auflösung von Objektreferenzen ebenfalls der Behebung von Synonymien. So kann z.B. eine Tasse sowohl als Becher als auch als Tasse bezeichnet werden (schwache Synonymie), trotzdem werden beide auf eine Objekt-ID abgebildet.

Suboptimale Segmentierung Im Rahmen von der thematischen Analyse von z.B. Nachrichtenkorpora stellt sich die Frage der Segmentierung nur eingeschränkt. So wird üblicherweise davon ausgegangen, dass ein Nachrichtenartikel genau eine Thematik behandelt. Problematischer ist z.B. die Segmentierung von Radionachrichten, deren Unterteilung in Nachrichtenbeiträge potentiell fehleranfällig ist.

Die Segmentierung von Dialogen in thematische Abschnitte muss üblicherweise ebenfalls anhand von Hinweisen geschehen, die nur mittelbar einen Themenwechsel anzeigen und die somit fehleranfällig sind. Fehlerhafte Segmentierung führt zu einer „Verrauschung“ der Statistik der Wort-Kookkurrenzen, die zu einer fehlerhaften Projektion der Wortvektoren in einen semantischen Raum führt. Dies wird deutlich an den Extrembeispielen der vollständigen Segmentierung, bei der jeder Kontext genau ein Wort enthält, oder dem der Bildung eines einzigen Kontextes, der alle Wortsymbole beinhaltet. Es ist nicht davon auszugehen, dass die in diesen Segmenten vorhandenen (bzw. nicht vorhandenen) Kookkurrenzinformationen für eine Themenerkennung verwendbar sind.

Um dem Problem der fehlerhaften Segmentierung zu begegnen, müssen unter Berücksichtigung der gegebenen Datenlage möglichst optimale Methoden zur Segmentierung gefunden werden. Problematischerweise variieren diese Methoden stark in Abhängigkeit von der Art der zu untersuchenden Kommunikation (Shriberg u. a., 2000). Die im Rahmen dieser Arbeit gewählten Segmentierungsverfahren werden, da sie auch in Abhängigkeit von dem Typ der Datenbasis variieren, an späterer Stelle beschrieben.

Zum Abschluss dieses Exkurses möchte ich noch kurz auf ein übliches Verfahren der Selektion relevanter Worte eingehen, der Gewichtung von Wortverteilungen.

Gewichte Bei dem einfachen Vektorraummodell, aber auch bei vielen der nachfolgenden Modelle ist es üblich, die relativen Häufigkeiten der einzelnen Vektoren für den jeweiligen Kontext zu gewichten. Eine sehr bekannte Gewichtung, deren Effektivität in Bezug auf *information retrieval* u.a. in (Spärck Jones, 1972), aber auch in Bezug auf Themenerkennungsaufgaben (Schultz und Liberman, 1999) bewiesen wurde, ist

die so genannte *inverse document frequency* (Salton und McGill, 1983):

$$idf = log(\frac{N}{n}) + 1 \tag{3.31}$$

wobei N die Anzahl der Dokumente und n die Anzahl der Dokumente ist, in denen das Wort (Symbol) vorkommt.

Der Grundgedanke der Gewichtung besteht in einer Beobachtung von Luhn (Luhn, 1958), nach der sowohl sehr selten vorkommende Worte als auch sehr häufig vorkommende Worte geringe Aussagen über die Textsorte (das vorkommende Thema, etc.) ermöglichen. Betrachtet man eine Häufigkeitsverteilung der nach Rang sortierten Worte, so stellt man fest, dass diese oft nach bestimmten Kriterien verläuft. Ein früher, aber bedeutender Versuch die zugrundeliegenden Verteilungsprinzipien zu erfassen, ist das so genannte **Zipfsche Gesetz** (*Zipfs Law*) (Zipf, 1949).

Demzufolge ergibt sich

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s} \tag{3.32}$$

N ist dabei die Anzahl der Elemente (Worte in dem Korpus), k der Rang und s ein Exponent, der häufig als 1 angenommen wird.

Multipliziert man die Häufigkeiten der einzelnen Symbole mit der *idf*, so ergibt sich eine Verteilung, bei der die sehr häufig vorkommenden Worte durch die Logarithmierung, die selten vorkommenden durch ihre geringe Vorkommenshäufigkeit benachteiligt sind. Auf diese Weise lässt sich auf einfache Weise das relevante Vokabular selegieren.

Problematischerweise schlägt aufgrund der kurzen Trainingssegmente bei der Themenerkennung auf HRI die Herausfilterung von selten vorkommenden Worten durch ihre geringe Vorkommenshäufigkeit fehl. Aus diesem Grund wird wie auf S.62 beschrieben im Kontext dieser Arbeit Gebrauch von einer minimalen Auftretenshäufigkeit C gemacht.

3.8 Der Sprung in die Multimodalität

Semantische Räume wurden in der Vergangenheit fast ausschließlich zur Verarbeitung von Daten auf rein textueller Basis verwendet. Da es ein erklärtes Ziel dieser Arbeit ist, die Fülle an multimodaler Information, die einem Roboter während einer HRI zur Verfügung steht, zu erschließen, stellt sich die Frage wie dies im Rahmen des dargelegten Verfahrens geschehen kann.

Grundsätzlich stehen zwei Ansatzpunkte zur Verfügung, an denen semantische Räume um multimodale Information bereichert werden können. Die erste Modifikation ist die Optimierung des Segmentierungsprozesses. Es ist zu erwarten, dass die Segmentierung unter Berücksichtigung multimodaler Hinweise darauf, an welchen Stellen ein (thematisches) Segment endet oder beginnt, wesentlich verbessert bzw. überhaupt ermöglicht wird. Auf die jeweiligen Segmentierungsprozesse wird an späterer Stelle eingegangen, daher möchte ich jetzt zum zweiten Punkt kommen.

Die zweite Alternative besteht in der Modifikation des Verfahrens selbst, so dass die direkte Verarbeitung multimodaler Information – im Gegensatz zu einem rein unimodalen, wortbasierten Ansatz – ermöglicht wird. Auf Vorarbeiten zu diesem Thema möchte ich im folgenden Abschnitt eingehen.

3.8.1 Vorarbeiten – SLSA und FLSA

Insbesondere mit zwei Verfahren wurde in der Vergangenheit versucht, die Performanz der LSA durch die Anreicherung mit zusätzlichen Informationen alternativer Modalitäten zu verbessern. Das eine Verfahren – **SLSA** (*Structured Latent Semantic Analysis*), dargestellt in (Wiemer-Hastings, 2000) und (Wiemer-Hastings und Zipitria, 2001), befasst sich insbesondere mit der Anreicherung der LSA durch syntaktische Informationen, die aufgrund des *bag of words*-Ansatzes und Prozessen wie der Lemmatisierung einer üblichen LSA nicht zur Verfügung stehen.

Die SLSA beruht im Wesentlichen auf drei Vorverarbeitungsschritten (Wiemer-Hastings und Zipitria, 2001):

- Der Auflösung von Anaphern; so z.B. der Ersetzung von Personalpronomina durch ihr Antezedens.
- Der Unterteilung von komplexen Sätzen in einfache Sätze.
- Der Segmentierung von Sätzen in Subjekt, Verb und Objektphrasen.

Tests haben ergeben, dass die SLSA unter bestimmten Voraussetzungen bessere Ergebnisse als eine unmodifizierte LSA erzielt. Auf der anderen Seite scheint die Effektivität des Verfahrens stark von den Umständen abzuhängen – so konnten z.B. in (Wiemer-Hastings, 2000) keine besseren, sondern eher schlechtere Ergebnisse als mit der einfachen LSA erzielt werden.

Im Rahmen dieser Arbeit wurden die vorgeschlagenen Modifikationen an der Segmentierung nicht übernommen. Dies liegt zum einen daran, dass die grammatische Struktur gesprochener Sprache nicht trivial zu ermitteln ist, zum anderen daran, dass das verwendete Verfahren auf das Lernen von Themen abzielt, wofür die thematische Segmentierung die relevanten Hinweise liefert. Eine solche Segmentierung ist aber nicht in Einklang zu bringen mit der grammatischen Segmentierung der SLSA.

Die Idee, anaphorische Relationen – insbesondere Personalpronomina durch Substitution mit dem Antezedens – aufzulösen, wurde jedoch im Rahmen dieser Arbeit übernommen. Sowohl die *offline*-Evaluationen auf dem Korpus als auch die *online*-Benutzerstudien verfügen über Mechanismen, die diesen Prozess ausführen.

Eine weitere, grundsätzliche Erweiterung der LSA wird ebenfalls in (Wiemer-Hastings und Zipitria, 2001) angesprochen: Der Einführung neuer Symbole, die – in diesem Fall – grammatische Information mit sich tragen. Das auf diese Weise modifizierte Verfahren heißt „**tagged LSA**“, da der Wortstrom mit den Ergebnissen eines automatischen Taggers angereichert wird. Auf diese Weise werden Worte um grammatische Information ergänzt, z.B. wird der englische Artikel „*the*“ durch das Suffix `_DT` für „Determinator“ ergänzt. Allerdings führte diese Modifikation in den durchgeführten Versuchen zu keiner Verbesserung, sondern nur einer Verschlechterung der Ergebnisse.

Ein sehr verwandtes Verfahren ist die **Feature Latent Semantic Analysis** (FLSA) (Serafin und Eugenio, 2004) (Serafin, 2003). Generisch betrachtet erlaubt dieses Verfahren die Erweiterung der LSA um *features*, die endlich viele, konkrete Werte annehmen können. Dazu wird die Term-Dokument-Matrix um weitere Dokumentvektoren erweitert, die für die jeweiligen, durch die *features* repräsentierten Eigenschaften stehen. Ein Beispiel aus (Serafin und Eugenio, 2004) ist die Erweiterung um Sprecherinformation. In einem aufgabenorientierten Diskurs, in dem es

einen „*giver*“ (Instrukteur) und einen „*follower*“ (Ausführenden) gibt, würden zwei neue „Wort“-Vektoren mit diesen Namen[29] eingeführt werden. Die „Häufigkeit“ kann entweder den Wert 1 (*feature* trifft zu) oder 0 (*feature* trifft nicht zu) annehmen.

Diese Erweiterung entspricht im Wesentlichen der Aufnahme eines neuen Symbols – nämlich des Featurenamens – in die Trainingsdaten und ist formal somit nur unwesentlich zu unterscheiden von der *tagged LSA* in (Wiemer-Hastings und Zipitria, 2001). Eine Erkenntnis von Serafin und Di Eugenio besteht wiederum in dem großen Einfluss, den die Wahl der korrekten *features* auf die Effizienz der FLSA hat. Sie untersuchen verschiedene Modelle, die eine optimale Wahl diesbezüglich ermöglichen soll.

3.8.2 Adaption

In Analogie zu den dargestellten Vorarbeiten wurden verschiedene Erweiterungen für diese Arbeit übernommen. Zum einen – wie im vorigen Abschnitt beschrieben – die Auflösung von anaphorischen Beziehungen durch Substitution der Personalpronomina durch das jeweilige Antezedens. Zum anderen stellte sich die Frage, welche *features* idealerweise gewählt werden sollten, um eine angepasste FLSA durchführen zu können. Im Gegensatz zu den Vorarbeiten wurde dabei auch versucht, das einfache Vektorraummodell so wie den Fuzzy-Semantics-Raum mit einer erweiterten Term-Dokument-Matrix zu erzeugen. Dies geschah, da kein offensichtlicher Grund einen Nutzen durch diese Erweiterung auf die LSA beschränkte.

Ein wesentlicher Vorteil, den ein situiert kommunizierender Roboter gegenüber einem auf Textdokumente arbeitenden Themenerkennungssystem hat, ist die Kenntnis der Situation. Themen in situierter Kommunikation beziehen sich oftmals auf Objekte der Situation. Aus diesem Grund wurde die Einbeziehung von Objektinformation zur Erweiterung der einfachen, wortbasierten semantischen Räume in Angriff genommen. Dazu wurden neue „Worte“, die jeweils eindeutig erwähnte Objekte kennzeichneten (IDs), in die Trainingsmenge aufgenommen. Diese Vorgehensweise unterscheidet sich von der FLSA in zwei Punkten:

- Das *feature*-Set ist nicht begrenzt, da eine beliebige Menge an Objekten an einer Situation beteiligt sein kann.
- Die Objekt-IDs **ersetzen** die ursprünglichen Bezeichner.

Dass nicht nur eine einfache Erweiterung um die Objekt-IDs vorgenommen wurde, sondern die referenzierenden Nominalphrasen durch die IDs ersetzt wurden, hat folgenden Grund:

Angenommen, in einer Situation bestehen zwei Themenbereiche – Essenszubereitung und Pflege von Topfpflanzen. In Kontext beider Themen wird je ein Messer erwähnt – das eine dient der Entfernung alter Pflanzenteile, das andere der Zerkleinerung von Nahrungsmitteln.

Wenn beide Objekte als „Messer“ bezeichnet werden, kann eine herkömmliche LSA nicht zwischen den Objekten differenzieren. Ein Clustern nach Themen würde somit Schwierigkeiten bei der Differenzierung der Themen haben. Dieser Effekt wäre immer

[29] Üblicherweise wird eine Differenzierung zwischen Worten und *features* eingeführt, so könnten letztere z.B. durch spitze Klammern gekennzeichnet werden.

noch – wenn auch abgeschwächt – gegeben, wenn sowohl die Objekt-IDs als auch die Originalbezeichnung „Messer“ in der Trainingsmenge vorhanden wären. Erst die Löschung des eine hohe Polytextie aufweisenden Symbols „Messer“ unterbindet diesen Effekt vollständig.

Zu beachten ist, dass die Auflösung von Objektreferenzen nicht nur der Löschung von Symbolen mit hoher Polytextie dient. Auch der schädliche Effekt von synonymen Objektreferenzen wird durch sie aufgehoben. Auf diese Weise ist zu erwarten, dass zumindest in den Fällen, in denen einzelne Objekte stets einzelnen Themen zugeordnet werden können, die Qualität des semantischen Raumes als Grundlage für einen Themenerkennungsvorgang stark durch diesen Prozess verbessert wird.

Bevor ich die endgültige Struktur des *offline*-Themenerkennungssystems anhand dieser Überlegungen skizzieren möchte, ist eine Anpassung der bisherigen Definition des Begriffes „Kommunikationssegment“ notwendig.

Definition 3.10 (Kommunikationssegment, multimodal) *Ein **multimodales Kommunikationssegment** besteht aus allen im Rahmen einer Kommunikation direkt oder indirekt produzierten Zeichen,*

*(i) die von **einem** Kommunikationsteilnehmer geäußert/produziert wurden,*

(ii) die zwischen zwei definierten Zeitpunkten t_1 und t_2 geäußert (produziert) wurden und

(iii) die einem oder mehreren definierten Typen angehören.

Der oder die Typen der Zeichen sind dabei je nach (Verwendungs-)Kontext frei wählbar.

Ein Zeichen wird direkt produziert, wenn es der Kommunikationsteilnehmer z.B. ausspricht oder gestisch ausdrückt. Ein Zeichen wird indirekt produziert, wenn es anhand eines definierten deterministischen Prozesses aus der Kommunikation ableitbar ist. So kann z.B. die (direkte) Äußerung des Wortes „Orange“ ebenso wie die (direkte) Äußerung des Wortes „Apfelsine“ das semantische Konzept „Zitrusfrucht“ (indirekt) instanziieren, obwohl dieses als Zeichen nicht geäußert wurde.

Mit Hilfe dieser Definition können auch Objektreferenzen o.ä. Teil eines Kommunikationssegmentes sein, so dass auch bei der multimodalen Adaption des Themenerkennungssystems weiterhin von Kommunikationssegmenten die Rede sein kann. Analog ist der Begriff *chunk* zu interpretieren.

3.9 Endgültige Struktur des offline-Themenerkennungssystems

Die in diesem Kapitel angeführten Überlegungen führten zu der endgültigen (internen) Struktur des *offline*-Themenerkennungsprozesses. Ich möchte sie an dieser Stelle noch einmal kurz zusammenfassen. Die Erweiterung des Prozesses für ein System, welches in Echtzeit auf einem Robotersystem arbeitet, wird in Kapitel 6 dargestellt.

Grundsätzlich besteht ein Themenerkennungsvorgang aus drei Schritten, nämlich:

- der **Datensammlung**. Dieser Schritt dient der Bereitstellung der Datenbasis, die zum Training verwendet werden kann und hängt ggf. stark vom Anwendungsgebiet ab. Insbesondere die Anreicherung mit multimodaler Information geschieht an dieser Stelle.

- dem **Training**, in dem aus der Datenbasis und unter Berücksichtigung von Verlaufsinformation (*history*) Themen identifiziert werden.

- dem **Tracking**, welches neue Dokumente/Äußerungen/etc. den durch das Training erkannten Themen zuordnet.

Das Tracking wurde bereits umfassend beschrieben, die Datensammlungsschritte sind von dem Anwendungsgebiet abhängig und werden bei den einzelnen Experimenten geschildert. Das Training gliedert sich wiederum in fünf Unterschritte:

1. **Vorverarbeitung**: In diesem Schritt werden Funktionswörter und ggf. zu selten vorkommende Wörter gelöscht. Außerdem kann eine Gewichtung mit der *idf* oder einem vergleichbaren Faktor geschehen.

2. **Segmentierung**: Anhand anwendungsspezifischer Informationen wird eine Segmentierung der Datenbasis in thematische Abschnitte vorgenommen.

3. Konstruktion des **semantischen Raumes**: Aus der Datenbasis wird eine Term-Dokument-Matrix generiert, die dann als Vorlage für die Konstruktion eines semantischen Raumes dient. Zur Konstruktion des semantischen Raumes gehört die Kalkulation der Distanzen zwischen den einzelnen Wortvektoren.

4. **Clustern**: Der semantische Raum wird anhand der berechneten Distanzen mit Hilfe eines Clusteralgorithmus unterteilt.

5. Erzeugung der **Themenmodelle**: Die Cluster werden als Grundlage für die Berechnung von Themenmodellen verwendet, die Informationen über die durchschnittlichen Distanzen aller Symbole zu jedem Thema besitzen. Sie sind die Grundlage für den Trackingvorgang, der asynchron zum Training arbeiten kann.

Der Informationsfluss wird zur besseren Veranschaulichung in Grafik 3.2 auf der nächsten Seite dargestellt. Der an dieser Stelle dargestellte Aufbau des Systems kann bezüglich der Reihenfolge im Detail von der später vorgestellten Realisierung abweichen, so z.B. an der Stelle, an der die multimodale Erweiterung der Datengrundlage geschieht. Der Aufbau der Prototypensysteme folgt jedoch grundsätzlich dieser Struktur.

Nachdem die grundlegenden Elemente des Themenerkennungssystems feststanden, wurde ein Korpus aus multimodaler HRI aufgezeichnet, um auf diese Weise über eine Grundlage für anschließende Evaluationen zu verfügen. Das Korpus und die Erstellungskriterien werden im Folgenden dargestellt.

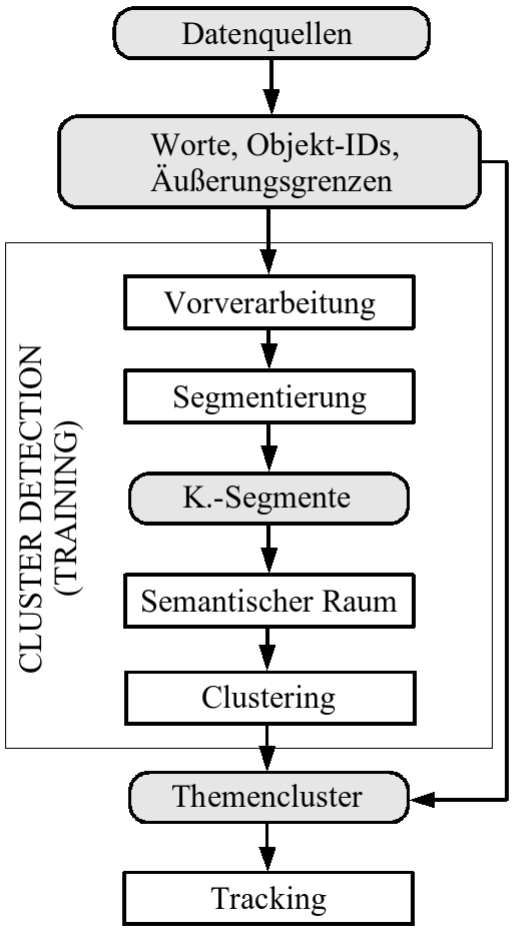

Abbildung 3.2: Basisstruktur der Themenerkennung

4 Korpus

Um eine solide Grundlage für die Entwicklung und die Evaluation von themenerkennenden Methoden in situierter Mensch-Roboter-Kommunikation zu schaffen, wurde in einem frühen Stadium des Promotionsvorhabens ein Datenkorpus erstellt (Maas und Wrede, 2006). Das Korpus wird im Folgenden als **BITT-Korpus** (Bielefeld Topic Tracking) bezeichnet. Trotz des großen Aufwands, den die Erstellung eines solchen Korpus mit sich brachte, erwies sie sich als notwendig, da bestehende Korpora, die situierte, multimodale Kommunikation zum Inhalt haben, nicht die für eine Themenerkennung unter den geforderten Umständen relevanten und interessanten Eigenschaften besaßen.

So existieren z.B. im Kontext des Sonderforschungsbereiches 360 „Situierte Künstliche Kommunikatoren“ aufwändig annotierte Dialoge, in denen zwei Menschen miteinander kommunizieren um ein Spielzeugflugzeug zu konstruieren. In diesen Dialogen nimmt eine Versuchsperson die Rolle des Instrukteurs ein, dem entweder eine Anleitung vorliegt, wie das Flugzeug zu konstruieren ist, oder aber ein Modell des fertigen Flugzeugs selbst. Dem Konstrukteur liegen die einzelnen Bauteile des Flugzeugs vor. Instrukteur und Konstrukteur sind in (fast) allen Dialogen durch einen Sichtschirm voneinander getrennt.

Trotz der offensichtlichen Situiertheit der Dialoge, die in den meisten verfügbaren Korpora nicht gegeben ist, hielten mich im Wesentlichen drei Gründe von der Verwendung dieses Korpus ab:

1. Die thematische Vielfalt dieser Dialoge ist sehr gering: Zwar existieren bisweilen Subdialoge, in denen spezielle Elemente des Flugzeugs – wie z.B. Heckflosse, Fahrwerk etc. – zusammengebaut werden, allerdings finden sich diese nicht in jedem Dialog in vergleichbarer Form wieder. Auf diese Weise entsteht eine Landschaft von Dialogen, in denen Themenkennung als außerordentlich schwierig zu beurteilen ist. Da schon andere, selbst auferlegte Beschränkungen wie dynamische Themen, natürliche Kommunikation etc. eine Themenerkennung stark erschweren, stellt dieses Korpus alles andere als eine optimale Grundlage zum Testen neuer Algorithmen dar.

2. Die Kommunikation zwischen einem künstlichen Kommunikator (d.h. Roboter) und einem natürlichen Kommunikator würde mit großer Wahrscheinlichkeit andere Formen annehmen als die Kommunikation zwischen zwei Menschen. Zwar soll das zu erstellende themenerkennende System in der Lage sein, Themen in möglichst natürlicher Mensch-Roboter-Kommunikation zu erkennen, welche mit reiner Mensch-Mensch-Kommunikation im Idealfall in vielen Aspekten identisch sein soll, aber die Wahrscheinlichkeit, dass in der Praxis starke Unterschiede in den Kommunikationsformen existieren, ist sehr groß. Um ein Beispiel zu nennen: In einer Mensch-Mensch-Kommunikation können feine Unterschiede in der Prosodie Änderungen in der Bedeutung einer Äußerung bewirken. Sollte

eine Maschine nicht in der Lage sein, diese Änderungen korrekt zu interpretieren, wird sich der Mensch mit großer Wahrscheinlichkeit an diese Situation gewöhnen und Prosodieänderungen nicht mehr in der fehlgeschlagenen Weise einsetzen. Enthält aber ein Testkorpus für eine Themenerkennung solche Hinweise, die dann für das Themenerkennungssystem utilitarisiert werden, dann wird es in Kombination mit der Maschine wahrscheinlich Fehler erzeugen.

3. Es ist anzunehmen, dass die thematische Struktur eines situierten, aufgabenorientierten Dialogs sich nicht ausschließlich an allgemeinen Diskursprinzipien, sondern auch an den Gegebenheiten der Situation und der Aufgabenstellung orientiert. Aus diesem Grund sollte der für die Entwicklung der Themenerkennung zu verwendende Dialog von der Situation und Aufgabenstellung möglichst nahe an dem potentiellen Anwendungsgebiet der Themenerkennung liegen. Da weiterhin davon auszugehen ist, dass eine automatische Themenerkennung über Parameter verfügt, die für die jeweilige Aufgabenstellung angepasst werden müssen, sollten die Parameter schon anhand des Entwicklungskorpus abschätzbar sein.

Ein weiteres Korpus, das im Rahmen des SFB-360 erstellt wurde – das so genannte „Wizard-of-Oz"-Korpus (Brindöpke u. a., 1997) – kompensiert zumindest den zweiten Aspekt: In diesem Korpus wird die Rolle des Konstrukteurs zwar von einem Menschen übernommen, allerdings wird der Instrukteur in der Überzeugung gelassen, dass es sich bei dem für ihn unsichtbaren Konstrukteur um ein künstliches kommunizierendes System handelt. Dieser Typ von Korpora/Experimenten wird wie oben in Anlehnung an das bekannte Kinderbuch (Baum, 1900) als „Wizard-of-Oz"-Experiment (bzw. -Korpus) bezeichnet. Die Kommunikation des „Wizards", also des falschen künstlichen Kommunikators, geschieht meist durch ein standardisiertes *interface* aus fest vorgegebenen Antwortmöglichkeiten, um die Glaubwürdigkeit der Täuschung zu untermauern und die Antwortvielfalt auf ein überschaubares Maß zu begrenzen.

Auch dieses Korpus erwies sich aus dem o.g. ersten und dritten Grund als weitgehend ungeeignet, auch wenn ein gut funktionierendes System zur automatischen Erkennung von Themen in Dialogen auch auf diesem Korpus akzeptable Ergebnisse erzielen sollte.

Nach eingehender Betrachtung der o.g. Korpora und der Erwägung der Verwendung weiterer, auf deren Problematiken ich im Weiteren nicht eingehen werde, stellte sich immer deutlicher heraus, dass die Erstellung eines eigenen Korpus notwendig sein würde.

4.1 Gestaltungskriterien

Ich werde im Folgenden detailliert auf die Gestaltungskriterien eingehen, die der endgültigen Fassung des Korpus zugrundelagen.

Situative Angemessenheit Da sich die Art der Kommunikation – und somit auch die Modalitäten der thematischen Strukturierung – stark durch die gegebene Situa-

tion ändern können, war es notwendig, Vorgaben zu der Situation zu machen[1]. Die im Rahmen des Forschungsprojekts COGNIRON stattfindende Forschung zielt wie geschildert auf die Entwicklung eines Haushaltsroboters bzw. Roboter-Companions hin; es wurden so genannte Basisszenarien entwickelt, die Meilensteine für die Entwicklung eines solchen Roboters darstellen. Sie sind zur Erinnerung an dieser Stelle noch einmal aufgeführt:

1. *robot home tour*: Der Roboter soll bei einer von einem Menschen geleiteten Tour durch eine Wohnung möglichst viel über diese lernen können.

2. *curious robot*: Der Roboter soll proaktiv möglichst viel über seine Umgebung erfahren, wobei Erkennung von Objekten und selbst initiierte Kommunikation mit Menschen im Vordergrund stehen.

3. *learning skills and tasks*: Der Roboter soll in Kommunikation mit dem Menschen fähig sein, neue Aufgaben und Fertigkeiten zu erlernen.

Als Basissituation für das Korpus wurde das *robot home tour*-Szenario gewählt. Es bietet den Vorteil einer großen thematischen Vielfalt, für deren Strukturierung der menschliche Kommunikationspartner Sorge tragen muss und die somit von dem Roboter erkennbar/erlernbar ist. Themenwechsel werden ausschließlich von den menschlichen Kommunikationspartnern initiiert. Im Falle z.B. eines „neugierigen Roboters“ (*curious robot*) aber würden nahezu alle Themenwechsel von dem Roboter initiiert werden, weswegen eine Hauptaufgabe einer Themenerkennung, nämlich die Erkennung von Themengrenzen, trivial sein würde. Meiner Ansicht nach stellt somit die Lösung des Problems der Themenerkennung für ein *home tour*-Szenario auch die zumindest theoretische Lösung desselben Problems für ein *curious robot*-Szenario dar, aber nicht *vice versa*.

Ähnlich verhält es sich mit dem Szenario „*learning skills and tasks*“. Sollte der Roboter durch explizite Benutzeraufforderung („jetzt kümmern wir uns um den Abwasch“) über die anstehende Aufgabe informiert werden, ist die Aufgabe der Themenwechselerkennung wiederum einfach zu lösen – es ist anzunehmen, dass eine zu erlernende Aufgabe mit einer Thematik übereinstimmt.

Die Wahl des *home tour*-Szenarios als Grundlage für die Erstellung des Korpus begünstigte auch die Entscheidung, den Roboter BIRON[2] in den Experimenten zu verwenden. Dies geschah unter anderem aus dem Grund, dass BIRON speziell für die Teilnahme an einem *home tour*-Szenario entwickelt wurde, und auf diese Weise nahezu optimale Voraussetzungen für die Experimente mit sich brachte. Im weiteren Verlauf dieser Arbeit wurde dann die Entscheidung gefällt, dass BIRON auch als Plattform für die weitergehende Entwicklung des Themenerkennungssystems fungieren sollte; insbesondere BIRONs Fähigkeiten zu komplexer, natürlichsprachlicher Interaktion – auch wenn sie für die Experimente nur äußerst eingeschränkt benötigt wurden – machten ihn zu einer idealen Grundlage für die Entwicklung und Integration eines Themenerkennungssystems.

[1] Natürlich wäre es wünschenswert gewesen, eine unter allen möglichen Umständen perfekt funktionierende Themenerkennung zu konstruieren, aber aufgrund der Komplexität der Aufgabenstellung ist dies ein nicht realistisches Ziel.

[2] vgl. Abschnitt 6.1

Angemessenheit der Modalität Das Korpus soll aus multimodaler HRI unter möglichst realistischen Umständen bestehen.

1. Die Mensch-Maschine-Kommunikation soll im Idealfall der im Anwendungsfall möglichen entsprechen.

2. Beeinflussungen der menschlichen Versuchsperson durch künstliche, äußere Einschränkungen sollten weitgehend ausgeschlossen werden.

Die Umsetzung des ersten Teils dieses Kriteriums ist nicht unkritisch: Wie oben beschrieben zwingt jedes Gerät, mit welchem eine Person in Interaktion tritt – so auch Lichtschalter, Autos oder Computer – dem Menschen Rahmenbedingungen der Kommunikation auf. Dies ist auch von jedem natürlichsprachlich kommunizierenden künstlichen Kommunikator zu erwarten, von dem in den Bereich der *science fiction* zu verbannenden Beispiel eines von einem Menschen nicht unterscheidbaren Androiden einmal abgesehen.

Das Ziel dieser Arbeit ist es, eine im Idealfall allgemein einsetzbare Methode zur Themenerkennung für Roboter in situierter Kommunikation zu finden, dabei dürfen aber Adaptionen an Situationen oder Hardware nicht von vorne herein ausgeschlossen werden. Da sich diese Arbeit wie beschrieben im konkreten Anwendungsfall mit der Themenerkennung auf der Roboterplattform BIRON beschäftigt, sollte das gewählte Korpus idealerweise aus typischen Dialogen von Menschen mit diesem Roboter bestehen, also aus Dialogen, wie sie im Anwendungsfall zustande kommen würden.

Dieses Ziel konnte aufgrund des momentanen Entwicklungsstadiums von BIRON sowie aus anderen Gründen nicht vollständig erfüllt werden. Ich fasse in der folgenden Aufzählung die Beschränkungen sowie die gewählten Kommunikationsmodalitäten von BIRON zusammen. Es ist zu erwarten, dass sich diese in näherer bis mittlerer Zukunft stark erweitern werden.

- Aus Gründen der Betriebssicherheit mussten die motorischen Fähigkeiten von BIRON auf ein Minimum reduziert werden, so dass der Roboter in den Experimenten nur in der Lage war, Kommunikationspartner zu erkennen und mit der Pan-Tilt-Kamera und durch Basisrotation zu verfolgen.

- Da die verwendete Dialogsoftware zu dem Zeitpunkt der Experimente nur einen Teil der für einen Haushaltsroboter interessanten Kommunikationsfunktionen unterstützte, blieb die Wahl zwischen (i) einer Erweiterung der Software, (ii) einem Wizard-of-Oz-Experiment und (iii) dem vollständigen Verzicht auf Sprachausgabe durch den Roboter. Die Modellierung eines Wizard-of-Oz-*interfaces* hätte sich als sehr komplex erwiesen, zusätzlich hätte mit einem solchen *interface* die Gefahr eines *biasing* der Versuchspersonen bestanden. Ziel war aber die Entwicklung eines Themenerkennungssystems (und damit eines Korpus) für im Idealfall maximal natürliche und damit unbeeinflusste bzw. uneingeschränkte Mensch-Roboter-Kommunikation. Die Gefahr des *biasing* stellte auch den Grund dar, aus dem auf eine Erweiterung des Dialogsystems verzichtet wurde, so dass die dritte Option – der vollständige Verzicht auf eine Sprachausgabe – gewählt wurde.

- Die Verwendung von BIRON als Plattform verhinderte ebenso den Ausdruck von Körpersprache und Gestik. Mimik wurde ausschließlich durch eine Ansteuerung des "Robotergesichts" simuliert: BIRON zeigte ein fröhliches Gesicht,

Abbildung 4.1: Teil des Experimentalraums

wenn er Spracheingabe von der Versuchsperson erkannte, ansonsten zeigte er ein neutrales Gesicht. Auf diese Weise sollte die Simulation von Aufmerksamkeit unterstützt werden.

Thematische Vielfalt Trivialerweise müssen die zum Testen verwendeten Datensätze auch über eine ausreichende thematische Vielfalt verfügen, um als Testdatensätze für eine Themenerkennung geeignet zu sein.

Dies wurde insbesondere durch eine reichhaltige Ausstattung des Experimentraumes mit diversen, aus verschiedenen Themengebieten stammenden Haushaltsgegenständen bewirkt. Dabei wurde darauf geachtet, die Gegenstände – wie es in einem realen Haushalt der Fall sein würde – thematisch zu gruppieren. Auf diese Weise entstanden ein Arbeitsplatz, ein Ort zum Teetrinken, eine (vorher schon existente, aber mit weiteren Objekten versehene) Küchenzeile etc. Ein Ausschnitt des voll ausgestatteten Experimentalraumes findet sich in der Abbildung 4.1.

Auf der linken Seite der Abbildung lässt sich ein Teil der Teeküche erkennen; weiterhin ein Sofa mit Stofftieren, ein Teetisch und ein Schränkchen mit einem Pokal, Süßigkeiten und Nähzeug. Diese Bereiche wie auch die Regale im Hintergrund wurden von den Versuchspersonen oft als eigene thematische Gebiete behandelt. Weiterhin mussten die Versuchspersonen angeregt werden, möglichst viel über die im Raum befindlichen Gegenstände zu erzählen. Dies geschah durch die Anweisung, BIRON genau über alles zu instruieren, da dieser in Zukunft Kinder im Alter von sieben Jahren

durch diesen Raum führen würde. Obgleich diese Anweisung nicht genau dem *home tour*-Szenario entspricht, stellte sich sich dennoch als effektives Mittel heraus, um die Kommunikationsbereitschaft der Versuchspersonen anzuregen. Alternativ hätte die Gefahr bestanden, dass die Versuchspersonen zu viel für BIRON zur Verfügung stehendes Weltwissen angenommen hätten, so dass nur minimale Erklärungen zu der Funktionalität der einzelnen Gegenstände abgegeben worden wären. Ausgiebige Erklärungen dienen aber wiederum der thematischen Vielfalt, so dass der oben genannte Trick angewandt wurde.

4.2 Ablauf der Experimente

Die Versuchspersonen wurden stets nach demselben Ablauf durch die Experimente geführt. Zuerst wurde die Versuchsperson nach ihrem Erscheinen in den Versuchsraum geführt, wo ihr ein Experimentbogen mit Anweisungen ausgehändigt wurde. Eine Kopie des Experimentbogens findet sich im Anhang dieser Dissertationsschrift (S.175f). Nachdem die Versuchsperson die Anweisungen gelesen hatte, wurden mögliche Fragen der Versuchsperson beantwortet, wobei keinerlei Fragen zu den Fähigkeiten des Robotersystems beantwortet wurden, um *biasing* zu vermeiden.

Anschließend wurde der Versuchsperson Zeit gegeben, sich mit dem Inhalt des Raumes vertraut zu machen. War diese Phase abgeschlossen, wurden die Aufnahmegeräte gestartet und die Versuchsperson aufgefordert, mit der Präsentation des Raumes zu beginnen. Diese Präsentation dauerte so lange, wie die Versuchsperson es für nötig befand, wodurch stark unterschiedliche Längen der einzelnen Monologe zustandekamen. Anschließend wurde die Aufzeichnung eingestellt und die Versuchsperson wurde gebeten, einen Bogen auszufüllen, in dem diverse Fragen gestellt wurden. Die Fragen unterteilten sich in drei Gebiete, zum einen allgemeine Fragen über die Person wie Alter und Studiengang, um eventuell auftretende Beeinflussungen der Experimente durch diese Daten rekonstruieren zu können. Weiterhin wurden Fragen gestellt, die für eine sich möglicherweise anschließende automatische Verarbeitung der Tondaten hätten als nützlich erweisen können, wie die Frage, ob die Person einen besonderen Dialekt besitzt, welches ihre Muttersprache ist und ob die Versuchsperson raucht. Ein dritter Fragenkatalog beschäftigte sich mit der Benutzerfreundlichkeit von BIRON, so wurden die Versuchspersonen aufgefordert, ihnen unangenehme oder angenehme kommunikative Eigenschaften des Roboters zu nennen bzw. die Versuchspersonen wurden gefragt, in wie weit sie sich in der Kommunikation mit einem menschlichen Kommunikationspartner anders verhalten hätten. Im Anhang findet sich eine Auflistung der den Versuchspersonen gestellten Fragen. Zu bemerken ist an dieser Stelle noch, dass die Experimente wie die Experiment- und Fragebögen vollständig in englischer Sprache gehalten wurden. Der Grund dafür findet sich zum einen in den Sprachfähigkeiten von BIRON, die weitestgehend für das Englische entwickelt wurden, so wie dem Umstand, dass das Korpus nach Fertigstellung für eine möglichst große Anzahl von potentiell interessierten Forschern zur Verfügung gestellt werden sollte.

4.3 Aufzeichnung und technische Grundlagen

Bei der Aufzeichnung der Experimente wurden zwei verschiedene Gruppen von Aufzeichnungsquellen verwendet. Auf der einen Seite war es notwendig, sämtliche Rohdaten der Robotersensoren, die für ein Themenerkennungssystem nützlich sein können, aufzuzeichnen. Auf der anderen Seite mussten die Experimente aus Perspektiven aufgezeichnet werden, die für die Sichtung und Nachbearbeitung des Materials durch Menschen besser geeignet sind. Folgende Aufzeichnungsgeräte und -quellen wurden verwendet:

1. Ein Camcorder mit Weitwinkelobjektiv, dessen Audioeingang an ein den Versuchspersonen aufgesetztes Headsetmikrophon angeschlossen wurde. Die durch diesen Recorder aufgezeichneten Videos lieferten die Grundlage für die im Folgenden beschriebene automatisch gestützte Nachbearbeitung der Experimentdaten.

2. Ein Rechner, der das Signal des Headsetmikrofons separat digital aufzeichnete. Auf diese Weise wurde eine bessere Tonqualität erzielt als durch die analoge Zwischenspeicherung auf dem Camcorder möglich gewesen wäre.

3. Ein Camcorder, der die Bilder der Pan-Tilt-Kamera des Roboters aufzeichnete, ohne Ton.

4. Ein Rechner, der über Firewire die Bilder der Stereokamera des Roboters aufzeichnete, ohne Ton.

5. Ein Rechner, der die Signale der Stereomikrophone des Roboters digital aufzeichnete.

Trotz der Vielzahl an Aufzeichnungsquellen und -geräten wurden wie oben angedeutet für die Weiterverarbeitung im Rahmen der Entwicklung eines automatischen Themenerkennungsmoduls für den Roboter nur die audio-video-synchronisierten Daten des Camcorders mit Weitwinkelobjektiv[3] verwendet, zumal sich die Qualität der Audiodaten auf diesem Camcorder für die halbautomatische Weiterverarbeitung als geeignet herausstellte.

4.4 Aufbearbeitung

Die Aufgabe der Nachbearbeitung der gewonnenen Audio- bzw. Videodaten bestand insbesondere darin, einem Themenerkennungsmodul in einer simulierten Roboterumgebung die Datenbasis zur Verfügung zu stellen, die dem Modul im Betrieb ebenfalls zur Verfügung stehen würde – also die durch den Roboter auszuführenden Vorverarbeitungsschritte zu simulieren. Um die Qualität der Themenerkennung zu evaluieren und zu verbessern, sollte diese Datenbasis daher möglichst optimal sein in der Hinsicht, dass sie von Fehlern, wie sie auf einem real funktionierenden Robotersystem auftreten würden, abstrahieren sollte. So sollten z.B. keine Fehler aus der Spracherkennung oder dem Personentracking in der Datenbank simuliert werden. Dies hat

[3] Im folgenden: Raumkamera

natürlich den Nachteil, dass das Themenerkennungsmodul in der Entwicklungsphase nicht auf Robustheit hinsichtlich dieser Probleme optimiert werden konnte. Auf der anderen Seite war davon auszugehen, dass keine spezielle Anpassung für Probleme dieser Art notwendig sein würde, sondern dass ein unter optimalen Umständen maximal gut funktionierendes Themenerkennungsmodul auch unter suboptimalen Umständen besser funktionieren würde als andere, zumal mögliche Fehlerquellen meist nur den Effekt der Verschlechterung des - im übertragenden Sinne gesprochen - Signal-Rausch-Abstandes der Themenerkennung zur Folge hätten.

Die Informationen, die das Themenerkennungsmodul benötigen würde, ließen sich in zwei Kategorien unterteilen:

- die Informationen, die für die automatische Segmentierung in Themen benötigt werden

- die Informationen, die für den Aufbau des semantischen Raumes benötigt werden (Symbolcluster)

Problematischerweise war abzusehen, dass die Frage nach der automatischen Segmentierung in thematische Symbolgruppen unterschiedliche Antworten auf dem Robotersystem und in den Experimentdaten verlangen würde. Dieser Umstand begründet sich insbesondere in zwei Umständen:

1. Monologstruktur der Experimente: Sehr hilfreiche Hinweise über Themensegmentierung in gesprochener Sprache finden sich z.B. in der Pausenstruktur der jeweiligen Kommunikation. So deuten z.B. in Nachrichtendaten Pausen von Nachrichtensprechern Themenwechsel an (Shriberg u. a., 2000). In den Experimentdaten findet sich dieser Effekt ebenfalls, da es sich bei den aufgezeichneten Experimenten um Monologe handelt. In Dialogsituationen jedoch kommen bestimmte Sprecherwechsel-Pausen hinzu, die insbesondere bei der Kommunikation mit einem Roboter sehr lang ausfallen können: Z.B. in Situationen, in denen der Benutzer ein „Habe verstanden“-Signal von dem Roboter erwartet. Da in HRI aufgrund der Störanfälligkeit typischer Spracherkennungssysteme häufig nach jedem Satz des Benutzers ein solches Signal von dem Roboter erwartet wird, sind Pausen in Mensch-Roboter-Dialogen nicht als Hinweise für Themenwechsel zu gebrauchen.

2. Bedauerlicherweise entfernten sich z.B. viele Versuchspersonen zu weit von dem Roboter, um den Robotersensoren verwertbare Audiodaten zu liefern, so dass die Aufnahme der Person durch die Robotersensoren in einigen Fällen nicht gewährleistet war. Diese Problematik entstand u.a. durch mangelndes Feedback des Roboters in den Experimenten. So war es z.B. nicht möglich, die auf dem Roboter vorhandenen Informationen über Blickkontakt aus den Experimentdaten zu erhalten. Aus diesem Grund konnten diese Informationen - anders als im Betrieb - nicht verwendet werden.

Natürlich wäre es möglich gewesen, auch die unter Punkt 2 genannten Daten durch manuelle Annotation in den Experimentkorpus mit aufzunehmen; auf eine solche Vorgehensweise wurde allerdings aufgrund des extremen Aufwands verzichtet, so dass das

Korpus nicht für die Evaluation der Segmentierungsalgorithmen, die später implementiert wurden, verwendet wurde. Für die Evaluation des Kernalgorithmus eignete sich das Korpus jedoch hervorragend.

Um einen Eindruck zu gewinnen, wie gut der Prozess der Themenerkennung bei idealer Segmentierung funktionieren würde, wurde weiterhin durch manuelle Annotation die Grundlage für eine solche Segmentierung geschaffen.

Die folgende Liste stellt eine Zusammenfassung der Informationen dar, die von Themenerkennungsystemen benötigt werden und somit von dem Korpus geliefert werden müssen. Dabei ist zu beachten, dass das Korpus zu einem Zeitpunkt entwickelt wurde, an dem unklar war, welche Informationen insbesondere für die Segmentierung exakt benötigt werden würden bzw. sich als verlässliche Merkmale erweisen würden. So fanden nach Vorexperimenten die F0-Verläufe keine weitere Verwendung, da sie sich als äußerst unzuverlässige Indikatoren für Themenwechsel herausstellten.

Wie beschrieben benötigt der Themenerkennungsalgorithmus Daten, um zwei verschiedene Aufgaben zu erfüllen: Segmentierung und Tracking. Segmentierung ist die Unterteilung von rezipierten Kommunikationen in thematische Abschnitte, um auf diese Weise den unüberwacht lernenden Algorithmus trainieren zu können. Indikatoren wie Pausen in Gesprächen o.ä. können auf Themenwechsel hinweisen und können somit für die Aufgabe der Segmentierung hinzugezogen werden. Informationen, die für das Tracking herangezogen werden, dienen der Wiedererkennung von bekannten Themen. So können z.B. thematisch geprägte Worte nützliche Indizien für ein Thema sein.

Prinzipiell lassen sich natürlich Hinweise, die für das Tracking relevant sind, ebenfalls für die Aufgabe der Segmentierung heranziehen. Tatsächlich handelt es sich um eine sehr übliche Vorgehensweise im Forschungsbereich der Textsegmentierung, z.B. Wortinformation für eine unüberwachte Segmentierung von Texten in thematische Abschnitte heranzuziehen (Hearst, 1997) (Bestgen, 2006) (Choi u. a., 2001) (Choi, 2000). Aufgrund der zu erwartend geringen Menge an Daten, die zum Training eines solchen Algorithmus zur Verfügung stehen würden, wurden im Rahmen dieser Arbeit jedoch größtenteils heuristisch zu erfassende Informationen für die Segmentierung herangezogen, so z.B. Informationen über den Dialogverlauf.

Die im Korpus erfassten Informationsquellen sind:

- F0-Verläufe (Segmentierung)
- Pausen (Segmentierung)
- (Lemmatisierte) transkribierte gesprochene Sprache (Tracking)
- Objekt-, Objektgruppen- und Objektklassenreferenzen (Tracking)

  sowie eine

- Manuelle Themenannotation (Evaluierung)

Wichtig ist, dass die ersten vier Informationsquellen dem Robotersystem grundsätzlich zur Verfügung stehen.

Die folgenden Abschnitte behandeln die Vorgehensweise bei der Aufarbeitung des Korpus im einzelnen.

4.4.1 Formanten

Als Grundlage sowohl für die Transkription (s.u.) also auch für die Analyse der Formanten wurde das Programm PRAAT verwendet (Boersma und Weenink, 2005) (Boersma und Weenink, 2001). PRAAT kann für jedes Frame einer Klangdatei den Wert des Basisformanten F0 schätzen und diesen in einem Textformat ausgeben. Für die im Rahmen dieser Arbeit unternommenen Voruntersuchungen dienten diese Dateien als Grundlage, da sie einfach von weiteren Programmen eingelesen und verarbeitet werden konnten. Dabei wurde auf eine exakte Übereinstimmung der Zeitstempel in diesen Dateien mit den Zeitstempeln in dem in einem XML-Format vorliegenden Korpus geachtet.

4.4.2 Pausen

Das Korpus wurde von mehreren Personen einheitlich mit Hilfe von PRAAT transkribiert. Als Grundlage dafür diente eine Vorverarbeitung, bei der Phasen mit Sprachaktivität von solchen ohne durch eine energiebasierte *voice activity dection* (VAD) unterschieden wurden. Dazu wurde dasselbe System[4] verwendet, welches auf BIRON Anwendung findet, um auf diese Weise Daten zu erhalten, die möglichst genau den auf dem Robotersystem zur Verfügung stehenden gleichen.

In der finalen XML-Variante des Korpus wurde jede Äußerung einer Versuchsperson mit einer Start- und Endzeit gekennzeichnet. Äußerungen sind dabei identisch mit den von dem VAD-System erkannten Zeitabschnitten. Abbildung 4.2 gibt ein Beispiel für eine solche Markierung.

<utterance start="35.11" end="36.39"> (...) ***and a lot of things*** *(...) </utterance>*
<utterance start="36.63" end="41.22"> (...) ***this desk for example*** *(...) </utterance>*

Abbildung 4.2: Pausendarstellung im Korpus

4.4.3 Transkription und Lemmatisierung

PRAAT diente ebenfalls als Grundlage für die manuell durchgeführte Transkription des Korpus.

Bei der Transkription handelt es sich im Wesentlichen um eine graphemische Transkription, bei der auf die Verwendung von Interpunktionszeichen verzichtet wurde. Für Hesitationen sowie für umgangssprachliche Worte mit mehreren oder keiner spezifischen Rechtschreibung (wie z.B. „OK“) wurden standardisierte Formen eingeführt.

Da angedacht war, den Korpus für das Training eines Spracherkennungssystems zu verwenden, wurden unüblich betonte Worte sowie Abbruchfehler speziell markiert, ebenso wie Störungen durch Hintergrundgeräusche. Für Störungen, Einschübe in deutscher Sprache, Anfragen an den Versuchsleiter bzw. Einwürfe des Versuchsleiters sowie Phasen unverständlicher Sprache von Seiten der Versuchsperson wurden

[4]Das VAD-System ist Teil des Toolkits ESMERALDA, vgl. (Fink, 1999).

XML-ähnliche Markierungen in Form von Tags eingefügt. Auf diese Weise konnte direkt aus der graphemischen Transkription die XML-Annotation, die dann schrittweise erweitert wurde, erzeugt werden.

Wie beschrieben ist die Umwandlung in Stammformen oder die Lemmatisierung ein üblicher Vorverarbeitungsschritt für Aufgaben im Bereich der Themenerkennung. Für den BITT-Korpus wurde eine automatische Lemmatisierung mit Hilfe eines statistisch arbeitenden Softwaretools, dem so genannten *Tree-Tagger* erzeugt (Schmid, 1995) (Schmid, 1994). In der finalen Version des Korpus befinden sich sowohl die lemmatisierten als auch die nicht lemmatisierten Formen der geäußerten Worte. Abbildung 4.3 zeigt die Verankerung dieser beiden Formen in der Annotation.

<utterance start="35.11" end="36.39"> <orig> ***I thought there were*** *(...) </orig> <lemma>* ***I think there be*** *(...) </lemma></utterance>*

Abbildung 4.3: Darstellung transkribierter Sprache im Korpus

Für die Themenerkennungsexperimente wurde ausschließlich die lemmatisierte Version des Korpus verwendet, allerdings war die nicht-lemmatisierte Form für manuelle Sichtungen des Korpus – z.B. im Rahmen der Annotation von Themen – von großem Nutzen.

4.4.4 Annotation von Referenzen

Für die Entwicklung des Themenerkennungssystems war es von besonderer Bedeutung, situierte Informationen zur Unterstützung mit einzubinden. Schon zu einem relativ frühen Punkt der Arbeit wurde vermutet, dass die Auflösung von Objektreferenzen wesentlich zu der Effizienz des Systems beitragen würde.

Die Auflösung von Objektreferenzen geschah manuell und ausschließlich für Objekte, die in der Situation gegeben waren. Dabei wurden sowohl Referenzen, die ausschließlich in sprachlicher Form gegeben waren, aufgelöst, also auch Referenzen, die durch eine Kombination aus Sprache und Gestik entstanden. Rein auf Gesten basierende Referenzen wären auch annotiert worden, allerdings waren diese im Korpus nur in einer vernachlässigbaren Anzahl vorhanden. Zu beachten ist, dass mit Hilfe semiautomatischer Prozesse sowohl die lemmatisierten als auch die nicht-lemmatisierten Teile des Korpus identisch annotiert wurden.

Aufgrund der Grammatik der englischen Sprache erwies sich die Annotation von Objektreferenzen als nichttriviale Aufgabe. Nominalphrasen können sowohl definit auf einzelne Objekte referieren („*Peter*“, „*the table*“), aber auch definit auf Gruppen („*the bon-bons*“). Ebenso können indefinite Objekte und Gruppen referenziert werden („*a mouse*“, „*mice*“). Weiterhin können Nominalphrasen in kopulativer Verwendung („*this is a table*“) eine Typzugehörigkeit ausdrücken.

Der Erstellung des Annotationsschemas lag weniger der Gedanke zugrunde, eine grammatisch ausgefeilte Annotation zu finden, sondern eine anwendungsorientierte. Dies geschah nicht zuletzt aus dem Gedanken, dass die Annotation das für das Robotersystem während der Kommunikation erhältliche Wissen umfassen sollte. Aus diesem Grund wurden im Rahmen der Annotation folgende Fälle unterschieden:

Referenzen auf Einzelobjekte: Referenzen auf Einzelobjekte wurden im Korpus durch XML-Elemente mit dem Namen „*object*“ gekennzeichnet. Jedes Objekt innerhalb des Versuchsraumes erhielt eine ID, auf die mit einem Attribut eines Kindelements referenziert wurde. Abbildung 4.4 zeigt die Annotation einer Referenz auf ein einzelnes Objekt, einen schwarzen Stoffraben. Zu beachten ist insbesondere das *pos*-Element. Üblicherweise umfasst es die komplette Nominalphrase, inklusive Determinator und Attribute. In bestimmten Fällen musste das Element jedoch in zwei Elemente aufgeteilt werden, da die Nominalphrase sich – durch eine Sprecherpause getrennt – über zwei *utterance* -Elemente erstreckt. In diesem Fall wurden sowohl der Beginn als auch das Ende der Nominalphrase als Referenz auf das entsprechende Objekt annotiert.

there is *<object><reference oid=”raven_01”/>*<pos> **a black raven** </pos> </object>(...)

Abbildung 4.4: Referenz auf Einzelobjekte

Die Annotation von Objekten mit IDs basiert auf einer großen Anzahl von nicht trivial zugänglichen Informationsquellen. So wird in bestimmten Fällen die Interpretation von Zeigegesten benötigt, in anderen die Auflösung von anaphorischen Beziehungen (Mitkov, 2002), z.B. im Fall vom Personalpronomen „*it*“ oder einfach im Fall der wiederholten Nennung eines im Fokus befindlichen Objektes. Trotzdem kann in all diesen Fällen ein an der Kommunikation beteiligtes Robotersystem – wenn auch mit einer gewissen Fehlerrate, die im Korpus nicht widergespiegelt ist – die Objekte identifizieren: Für die Auflösung anaphorischer Beziehungen existieren mittlerweile leistungsfähige Algorithmen (ebd.), weiterhin ist die Komplexität der aufzulösenden anaphorischen Beziehungen zumindest im BITT-Korpus verhältnismäßig gering, so dass mit einfachen Methoden gute Ergebnisse erzielt werden können, die noch ggf. durch Rückfragen optimierbar sind. Die Identifikation gestisch referenzierter Objekte ist auf BIRON implementiert, vgl. dazu (Haasch u. a., 2005).

Referenzen auf Gruppenobjekte: Im Versuchsraum existierten Gruppen von Objekten, deren Einzelobjekte nie als solche referenziert wurden. Ein Beispiel dafür sind Aufkleber auf einem Abziehbogen (vgl. Abbildung 4.5). Da sowohl ein Objekterkennungssystem diese als ein Objekt identifiziert hätte als auch von der Sprachverwendung her die Gruppe stets als eine einzelne Entität auftrat, wurden solche Objekte ähnlich den Einzelobjekten behandelt. Allerdings wurden für sie ein spezieller Elementtyp *objects*, so wie spezielle IDs verwendet, um Verwechselungen auszuschließen.

then there are *<objects><group oid=”g_stickers_01”/><pos>**little stickers*** *</pos></objects>*

Abbildung 4.5: Referenz auf Gruppenobjekte

Gruppenreferenzen Spontan im Rahmen der Kommunikation gebildete Gruppen, die aus sonst auch im Einzelnen referenzierten Objekten bestanden, wurden ebenfalls in einem *objects*-Element spezifiziert. Dabei wurde allerdings auf die Angabe

eines Gruppennamens verzichtet, stattdessen wurden die einzelnen Objekt-IDs der beteiligten Objekte annotiert. Referenzen auf echte Gruppen stellen ein theoretisches Problem dar, da nicht immer klar ist, ob ein Robotersystem in der Lage ist, alle Einzelobjekte zu erkennen. Oftmals werden diese Gruppen nämlich über Hintergrundwissen referenziert, welches einer Objekterkennungssoftware nicht unbedingt zur Verfügung steht. Ein Beispiel aus dem Korpus wäre die Gruppe aller Pflanzen in dem Versuchsraum, bei der die Versuchsperson davon ausgeht, dass der Roboter in der Lage ist, auf Anhieb alle vorhandenen Pflanzen zu erkennen. In Rahmen der Annotation wurde vereinfachend davon ausgegangen, dass der Roboter tatsächlich dazu in der Lage wäre – so hätte ein Robotersystem die Objekte der Gruppe z.B. durch Nachfragen identifizieren können.

then we've got *<objects><group oid=""/><member oid="ruler_01"/><member oid="compasses_01"/><pos>* ***drawing equipment*** *<pos></objects>* ***for geometry***

Abbildung 4.6: Referenz auf Gruppen

Referenzen auf abstrakte Gruppen Bei der kopulativen Verwendung von „*is*“ werden zwei Nominalphrasen – üblicherweise ein Demonstrativpronomen und eine indefinite Nominalphrase – miteinander verknüpft („*this is a table*“). Im Rahmen der Annotation wurde die demonstrative Nominalphrase als Referenz auf ein Einzelobjekt annotiert, die indefinite Nominalphrase als Referenz auf ein Gruppenobjekt. Diese wurde jedoch über die ID speziell als **abstrakte Gruppe** gekennzeichnet. Der Nutzen dieser Vorgehensweise bestand darin, dass in solchen – vom Roboter erkennbaren – Fällen thematische Bezüge über Objekttypen hergestellt werden können, die sonst bei der reinen Referenzierung über IDs verlorengehen. Bei diesen Annotationselementen handelt es sich um die einzigen Referenzannotationen von nicht im Raum befindlichen – da abstrakten – Objekten.

so <object><candidate oid="raven_01"/><pos>**this**</pos></object>**is** <group oid="c_birds"/><pos>**a bird** </pos></objects>

Abbildung 4.7: Annotation kopulativer Verwendung von „*is*“

Auf diese Weise wurde das gesamte Korpus annotiert. Zu beachten ist dabei, das sowohl die lemmatisierten als auch die nicht-lemmatisierten Inhalte des Korpus durch automatische Prozesse identisch annotiert wurden, so dass sich in beiden Textinhalten dieselben Referenzannotate finden lassen.

4.4.5 Themenannotation

Um eine Grundlage für die Evaluation zukünftiger Themenerkennungssysteme zu schaffen, wurde das gesamte Korpus von drei verschiedenen Personen anhand der in ihm vorkommenden Themen annotiert. Dabei wurde jedem der Annotatoren folgende Vorgehensweise vorgeschrieben:

1. Vertrautmachung mit dem Korpus (d.h. Sichtung der Videos)

2. Definition einer Liste von nicht-hierarchischen Themen

3. Annotation des Korpus

Der Umstand, dass die Annotatoren vor dem Annotationsprozess eine feste Liste von Themen definieren sollten, führte dazu, dass

a) fast ausschließlich globale Themen – also in vielen Dialogen wiederkehrende Themen – annotiert wurden

b) monologübergreifende Themen erkannt werden konnten

Punkt a) führte zu besserer Vergleichbarkeit mit den anschließenden Themenerkennungsexperimenten, da das Themenerkennungssystem in einem *jack-knife*-Verfahren trainiert wurde und somit globale Themen – und nicht lokale – erkannte. Punkt b) widerspiegelt die Aufgabe des Themenerkennungssystems, wiederkehrende Themen (im Kontrast zu lokalen) zu erkennen.

Die Annotatoren hatten die Aufgabe, jeder Äußerung[5] genau ein oder kein Thema aus der von ihnen vordefinierten Liste zuzuordnen. In der Annotation wurde für jedes *utterance*-Element für jeden Annotator ein Attribut eingeführt, welches als Wert das annotierte Thema besitzt. Ein Beispiel dafür kann in Abbildung 4.8 betrachtet werden (in der Abbildung wurden diverse Tags entfernt, um die Übersichtlichkeit zu gewährleisten). ta, tb und tc symbolisieren in der Annotation die drei unterschiedli-

<utterance start="144.15" end="149.98" ta="sofa" tb="" tc="couch">
*<lemma>**er there be ... something ... to sit on there be a er a little ... sofa**...</lemma>(...)</utterance>*

Abbildung 4.8: Themenannotation im Korpus

chen menschlichen Annotatoren. Das Beispiel ist folglich so zu lesen, dass Annotator ta die Äußerung mit dem Thema „sofa“, tb mit keinem Thema und tc mit dem Thema „couch“ annotierte.

Die vollständige DTD des Korpus findet sich im Anhang (Abschnitt 8.3).

4.5 Statistiken

Ich möchte in diesem Abschnitt noch einmal im Überblick die wichtigsten Daten des Korpus stichpunktartig zusammenfassen.

- Das Korpus besteht aus 29 transkribierten und annotierten Monologen so wie den Sensordaten des Roboters, wobei ausschließlich die Daten der Laser-Distanzmessung nicht mit aufgezeichnet wurden.

- Es nahmen insgesamt 29 Personen an den Experimenten teil, davon waren 24 weiblich und 5 männlich.

[5] d.h. jedem *utterance*-Element

- Von diesen Versuchspersonen waren 27 keine muttersprachlichen Sprecher der englischen Sprache, zwei stammten aus englischsprachigen Gebieten. Alle Versuchspersonen schätzen ihre Fähigkeit, Englisch zu sprechen/zu verstehen als gut bis exzellent ein.
- Im Korpus wurden 2620 Referenzen auf einzelne Objekte annotiert, so wie 1419 Referenzen auf Objektgruppen oder abstrakte Objekte.
- Das Korpus besteht aus 11209 automatisch erkannten Phasen von Sprachaktivität („`utterance`“). Nach der Entfernung von Geräuschen, unverständlichen Äußerungen etc. blieben noch 6989 Elemente übrig, die die Grundlage für die Themenerkennung und -annotation darstellten.
- Durchschnittlich 4900 Äußerungen wurden je Annotator mit einem Thema annotiert.
- Das Korpus steht seit der LREC 2006 der wissenschaftlichen Öffentlichkeit auf Anfrage kostenfrei zur Verfügung. Allerdings beschränkt sich die Weitergabe aus datenschutzrechtlichen Gründen nur auf die anonymisierten Transkriptionen so wie die Grundfrequenzdaten.

Das Korpus – das wie beschrieben aus immerhin knapp 300 Minuten aufgezeichneten Videodaten besteht – stellte eine verlässliche Grundlage für die Entwicklung und Evaluation eines Themenerkennungssystems dar. In dem folgenden Kapitel möchte ich auf die Evaluationsprozesse eingehen, denen das entwickelte System anhand des dargestellten Korpus sowie des Reuters-Korpus unterworfen wurde.

5 Offline-Evaluation

In dem folgenden Kapitel möchte ich die Experimente darstellen, anhand derer das Themenerkennungssystem *offline* evaluiert wurde. Im Rahmen von *offline*-Evaluationen lassen sich natürlich nicht alle Aspekte des Systems hinreichend erforschen, so z.B. das Laufzeitverhalten, die Effizienz der Segmentierung oder die sinnvolle Einbettung in das Robotersystem. Diese Aspekte werden in späteren Kapiteln eingehender behandelt. Der Nutzen der *offline*-Evaluation besteht im Besonderen in der Wiederholbarkeit der Experimente unter identischen Bedingungen, so dass sich die Auswirkungen bestimmter Parametersetzungen genau differenzieren lassen. Weiterhin konnte den *offline*-Evaluationen eine weitaus größere Datenmenge zugrundegelegt werden, als *online*-Benutzerstudien dies zulassen würden. Auf diese Weise konnte eine mittlere Trainingszeit simuliert und das Verhalten des Systems unter dieser Rahmenbedingung erforscht werden.

Die in Abschnitt 5.2 dargestellte Evaluation diente einer ersten Einschätzung der Qualität des verwendeten Verfahrens, insbesondere des Klassifikators. Dazu wurde ein nicht-situiertes Nachrichtenkorpus – das Reuters-21578-Korpus – herangezogen. Dieses Korpus hat sich in der Vergangenheit zu einem Standardkorpus für die Evaluation von Themenerkennungsverfahren entwickelt.

Im Anschluss wird eine umfangreiche Evaluation auf dem in Abschnitt 4 dargestellten Korpus diskutiert. Im Gegensatz zum Reuters-Korpus besteht dieses Korpus aus situierten Mensch-Roboter-Kommunikationsdaten, so dass das dargestellte Verfahren seine spezifischen Vorteile nutzen kann. Diese Evaluation soll insbesondere die Auswirkungen von algorithmus-spezifischen Anpassungen für situierte Umgebungen zeigen, aber auch darlegen, dass das Themenerkennungsverfahren im Kontext von größeren Trainingsgrundlagen funktioniert, als man üblicherweise bei Benutzerexperimenten aus der Kommunikation gewinnen kann.

Bevor mit der Darstellung der einzelnen Evaluationen begonnen wird, ist jedoch eine Diskussion der üblichen Evaluationsmaßstäbe sowie der spezifischen Probleme bei der Evaluation unüberwacht lernender Klassifikationsprozesse notwendig.

5.1 Evaluationsmaße

Im Bereich der Textklassifikation bzw. der Themenerkennung haben sich in der Vergangenheit verschiedene Evaluationsmaße (vgl. (Hotho u. a., 2005), (van Rijsbergen, 1979)) als nützlich erwiesen, obwohl weitere Entwicklungen nicht ausgeschlossen sind.

Grundsätzliche Basis der Evaluation einer Themenerkennung ist natürlich die Beantwortung der Frage, wie gut eine – als *ground truth* betrachtete – manuelle und eine automatische Klassifikation übereinstimmen.

Das intuitiv naheliegendste Maß ist **Akkuratheit** (*accuracy*). Es entspricht dem prozentualen Anteil an korrekt klassifizierten Dokumenten in Bezug auf das gesamte

Korpus. Bei dichotomen[1] Klassifikatoren tritt das Problem auf, dass die Akkuratheit über allen Klassifikatoren gemittelt werden muss, da z.B. mehrere dichotome Klassifikatoren ein Dokument als ihrem Thema zugehörig einordnen können. Alternativ werden die Ergebnisse der Einzelklassifikationen miteinander verrechnet, so dass eine eindeutige Klassifikation als Ergebnis erzielt wird. Der Vorteil von Akkuratheit liegt in der intuitiven Aussagekraft dieses Evaluationsmaßes. Aussagen wie „98% des Korpus wurden korrekt klassifiziert“ lassen direkt auf die Leistungsfähigkeit des Klassifikators schließen.

Trotz dieses offensichtlichen Vorteils wurde Akkuratheit mittlerweile in vielen Anwendungsbereichen von anderen Evaluationsmaßen verdrängt. Ein Grund dafür lässt sich u.a. an dem Reuters-Korpus erkennen: Ein Großteil der in diesem Korpus vorkommenden Dokumente wird nur durch wenige Themen abgedeckt. Ein (dichotomer) Klassifikator, der ein selten vorkommendes Thema erkennen soll, tut also gut daran, alle Dokumente der Rückweisungsklasse zuzuordnen – aufgrund der vielen „Korrekt Negativen“ wird er eine hohe Akkuratheit erzielen. Analog dazu würde ein nichtdichotomer Klassifikationsvorgang, der alle Dokumente der am häufigsten vorkommenden Klasse zuordnet, unter bestimmten Bedingungen ungerechtfertigt sehr gute Ergebnisse erzielen, die die Vergleichbarkeit mit alternativen Verfahren mindern.

Aus diesen Gründen werden häufig statt Akkuratheit die Maße **precision** und **recall** verwendet (Hotho u. a., 2005). Sie berechnen sich wie folgt:

$$precision = \frac{\#(relevant \cap retrieved)}{\#retrieved} \tag{5.1}$$

und

$$recall = \frac{\#(relevant \cap retrieved)}{\#relevant} \tag{5.2}$$

Relevant sind die Dokumente, die in den Referenzdaten[2] als dem Thema/der Klasse zugehörig ausgewiesen sind, *retrieved* sind die Dokumente, die der automatische Klassifikationsmechanismus dem Thema/der Klasse zugehörig befunden hat. *Precision* ist somit das Verhältnis von korrekt klassifizierten Dokumenten zu der Gesamtmenge von als dem Thema zugehörig klassifizierten Dokumenten. Ein niedriger *precision*-Wert tritt z.B. ein, wenn der Klassifikator viele *false positives* generiert.

Recall gibt Aufschluß über das Verhältnis der Zahl der korrekt klassifizierten Dokumente zu der Anzahl der tatsächlich zu der Klasse gehörigen Dokumente. Die Anzahl der inkorrekt klassifizierten Dokumente ist dabei irrelevant; ein dichotomer Klassifikator, der alle Dokumente als seiner Klasse zugehörig markiert, hat immer einen *recall*-Wert von eins. Im Fall eines zu „großzügig“ eingestellten dichotomen Klassifikators würde also zwar ein hoher *recall*-Wert vorliegen, aber ein sehr geringer für *precision*. Der umgekehrte Fall gilt für einen „geizig“ arbeitetenden Klassifikator.

Ein Problem, welches durch die Verwendung von zwei unterschiedlichen Maßen entsteht, ist, dass nicht ein, sondern zwei Werte die Qualität eines Klassifikationsprozesses angeben. Im Fall von dichotomen Klassifikatoren wird dieses Problem oft da-

[1] Die Bezeichnung „dichotom“ im Zusammenhang mit Klassifikatoren bedeutet im Rahmen dieser Arbeit, dass ein Klassifikator nur in der Lage ist zu entscheiden, ob (oder wie stark) zu klassifizierende Entitäten zu einer einzigen Klasse gehören. Somit wird von ihnen oftmals die „binäre“ bzw. dichotome Entscheidung der Klassenzugehörigkeit getroffen.

[2] Also z.B. einer manuellen Klassifikation von Testdaten

durch angegangen, dass versucht wird, den so genannten *break even*-Punkt zu ermitteln, an dem *precison* und *recall* identische Werte haben. Der Gedanke hinter dieser Vorgehensweise ist, dass dichotome Klassifikatoren üblicherweise einen einstellbaren Parameter (*classification treshold*) besitzen, der ihre Empfindlichkeit regelt. Durch sukzessive Annäherung oder einfache *brute force*-Berechnungen kann die Einstellung dieses Parameters gewonnen werden, an dem der *break even*-Punt erreicht ist. Auch ist es in der wissenschaftlichen Literatur nicht unüblich, Kurvenverläufe anstatt des *break even*-Punktes anzugeben.

Für Fälle, in denen dies nicht möglich oder nützlich ist, hat sich ein weiteres Maß bewährt. Bei dem so genannten **f-measure** (im Folgenden: f-Wert bzw. f) handelt es sich um eine Mittlung über *precision* und *recall* (van Rijsbergen, 1979). Es wird wie folgt berechnet:

$$f = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5.3}$$

Hierbei handelt es sich um eine harmonische Mittlung von *precision* und *recall* (van Rijsbergen, 1979) (Rennie, 2004). In der allgemeinen Form kann das Verhältnis gewichtet werden:

$$f_\alpha = \frac{(1 + \alpha) \cdot \text{precision} \cdot \text{recall}}{(\alpha \cdot \text{precision} + \text{recall})} \tag{5.4}$$

f ist somit identisch mit f_1.

f ermöglicht eine Mittlung über *precision* und *recall*, bei der im Vergleich zum einfachen arithmetischen Mittel schlechte Werte eines der beiden Teilwerte stärker gewichtet werden. f kann ebenfalls Werte von über 0 bis 1 annehmen. Auch im Rahmen dieser Arbeit wird f verwendet, um Vergleiche zwischen Parametrisierungen zu ermöglichen.

Eine weitere Problematik besteht – wie schon oben geschildert – in der Mittlung über den Einzelergebnissen der Klassifikatoren. Ein in der Vergangenheit nicht unübliches Verfahren zur Mittlung ist *microaveraging* (Yang, 1997). Dazu wird für jeden Themenklassifikator derselbe Wert für den *classification treshold* eingestellt. Anschließend werden tabellarisch die Anzahl der korrekt erkannten, fehlerhaft erkannten, korrekt nicht erkannten und fehlerhaft nicht erkannten Dokumente aufsummiert und auf diesen Werten *precision* und *recall* berechnet. Dieses Verfahren fand z.B. in (Joachims, 1998) Anwendung.

In der vorliegenden Arbeit wird im Gegensatz zu (Joachims, 1998) und vielen anderen Arbeiten kein dichotomer Klassifikator eingesetzt. Da das erkannte Thema dem in Bezug auf die enthaltenen Symbole durchschnittlich nahesten Thema entspricht[3], wird immer genau ein oder kein Thema gefunden. Es ist somit nicht möglich, mehrere Themen für ein Dokument bzw. für einen *chunk* zu erkennen oder die Erkennung einzelner Themen zu begünstigen oder zu behindern. Aus diesem Grund wurde auf die Berechnung eines *break even*-Punktes verzichtet; dennoch wurden die resultierenden Werte für *precision*, *recall* und f berechnet, was jedoch zu Einbußen im Bereich der Vergleichbarkeit zu alternativen Verfahren führt, die auf gemittelten Werten basieren.

[3] vgl. Abschnitt 3.6

Problematik bei der Evaluation unüberwacht lernender Themenerkennungsverfahren An dieser Stelle möchte ich auf eine besondere Problematik hinweisen, die sich bei der Evaluation nicht-dichotom klassifizierender, unüberwacht lernender Klassifikationsprozesse ergeben kann: Um zu bestimmen, ob ein Dokument korrekt klassifiziert wurde, muss bestimmt werden, welche automatisch erkannte Kategorie welcher Kategorie aus den Referenzdaten entspricht. Bei überwacht lernenden, dichotomen Klassifikatoren ist dies trivial, da *a priori* jedem Klassfikator eine Klasse aus der Trainingsmenge zugeordnet wird. Im Fall von sich dynamisch bildenden, automatischen Kategorien ist die Zuordnung dagegen problematisch.

Eine Vorgehensweise, die u.a. von (Seo und Sycara, 2004) angewandt wurde, besteht darin, jeder Kategorie aus den Referenzdaten die Kategorie aus den automatisch klassifizierten Daten zuzuweisen, die am häufigsten für die Referenzkategorie erkannt wurde. Auf diese Weise kann es jedoch passieren, dass zwei unterschiedlichen Referenzkategorien dieselbe automatisch erkannte Kategorie zugewiesen wird. Ein Beispiel für diese Vorgehensweise findet sich in Tabelle 5.1.

Manuelles Thema	# Äußer.	Anteile automatische Themen					zugeordnetes Thema	% Akkur.
		A	B	C	D	E		
arbeiten	10	0	2	7	0	1	**C**	70
essen	12	2	1	3	0	6	**E**	50
fernsehen	8	1	0	7	0	0	**C**	87.5
schlafen	14	2	2	0	9	1	**D**	64.3

Tabelle 5.1: Berechnung von Akkuratheit bei verteilungsgestützter Zuordnung automatisch erkannter Themen.

Im Falle einer Bestimmung der Akkuratheit kann dies zu ernsten Problemen führen, da in diesem Fall die Klassifikation aller Dokumente eines Korpus als derselben Kategorie zugehörig ein maximal gutes Ergebnis bedeuten würde, was jedoch nicht im Willen des Anwenders liegen kann.

Eine mögliche Vorgehensweise besteht darin sicherzustellen, dass zwei unterschiedliche Referenzkategorien stets unterschiedliche automatische Kategorien zugewiesen bekommen. Nach welchen Kriterien dies aber geschehen soll, ist schwierig zu beurteilen – schließlich haben beide Referenzkategorien zuerst einmal dasselbe „Anrecht“ auf die am stärksten korrespondierende automatische Kategorie. Im Rahmen dieser Arbeit wurden verschiedene Wege gewählt, um dieses Problem zu lösen. Detaillierte Angaben dazu finden sich in den Darstellungen der jeweiligen Evaluationen.

Im Falle der Bestimmung eines f-Wertes ist das Problem aufgrund der Definition des *precision*-Maßes kompensiert. Im beschriebenen Fall würde dieses Maß wesentlich schlechtere Ergebnisse liefern als bei einer 1:1-Abdeckung. Trotzdem sind Fälle, in denen eine automatisch gefundene Kategorie zwei oder mehr Referenzkategorien umfasst, nicht automatisch disqualifiziert. Dies ist insbesondere für das Gebiet der automatischen Themenerkennung von Nutzen, da leicht Fälle denkbar sind, in denen ein Themenerkennungsalgorithmus ein mehreren Referenzthemen übergeordnetes Thema findet oder *vice versa*. Eine Evaluation der f-Werte würde diese Fälle nicht automatisch als falsch bewerten, wie es im Rahmen einer modifizierten Berechnung der Akkuratheit möglicherweise geschehen wurde.

Im Folgenden werde ich die einzelnen Evaluationsvorgänge sowie die erzielten Ergebnisse darstellen.

5.2 Präliminare Evaluation auf dem Reuters-Korpus

Erste Experimente mit der Themenerkennungssoftware wurden auf dem Reuters-21578-Korpus unternommen. Das Korpus ist im Kontext wissenschaftlicher Forschung frei verfügbar[4] und besteht aus 21.578 Meldungen, die von der Nachrichtenagentur Reuters im Jahre 1987 veröffentlicht wurden.

Dem Korpus liegt eine SGML-Annotation zugrunde, die für die Artikel u.a. Autor, Datum, Überschrift und Textinhalt kennzeichnet. Weiterhin wurden Themen annotiert, wobei für jeden Artikel eine beliebige Anzahl Themen aus einer Liste von 135 Themen von den Annotatoren benannt werden durften.

Eine wichtige Grundlage der Evaluationen auf dem Reuters-Korpus stellen die „Splits“ dar. Dabei handelt es sich um festgelegte Unterteilungen des Korpus in Trainings-, Test- und ignorierte Dokumente, anhand derer Themenerkennungsverfahren vergleichbar getestet werden können. Die drei bekanntesten Splits sind Lewis (Lewis, 1992a) (Lewis, 1992b) (Lewis und Ringuette, 1994), ModApte (Apté u. a., 1994a) (Apté u. a., 1994b) und ModHayes (Hayes u. a., 1990) (Hayes und Weinstein, 1990). Für Forschungen im Bereich der Themenerkennung wird im Allgemeinen der ModApte („modified Apte“)-Split verwendet, der den Versuch einer Eingrenzung des Lewis-Splits auf Dokumente darstellt, die für Themenerkennungsevaluationen geeignet sind.

Die Evaluation des Themenerkennungsverfahrens geschah in grober Analogie zu den Untersuchungen, die in (Joachims, 1998) bzw. (Joachims, 1997) unternommen wurden. In diesen Artikeln analysiert Joachims die Effektivität verschiedener Klassifikatoren – darunter Standardverfahren wie den Bayes-Klassifikator – im Vergleich zu den von ihm in diesen Aufsätzen propagierten Support-Vector-Maschinen. Leider konnte aufgrund der Verschiedenartigkeit der jeweiligen Verfahren (Joachims/diese Arbeit) keine echte Vergleichbarkeit gewonnen werden. Gründe dafür werden weiter unten erläutert. Grundsätzlich aber unterliegen unüberwachte Lernverfahren gegenüber überwacht arbeitenden dem Problem, dass sie mit weniger Information auskommen müssen - eben nämlich der Information, welches Trainingsdokument zu welchem Thema/zu welcher Kategorie gehört. Weitere Probleme entstehen durch die angewandten Evaluationsmaße, die sich nicht auf den im Rahmen dieser Arbeit angewandten Klassifikationsmechanismus anwenden ließen. Trotzdem liefert (Joachims, 1998) gute Einblicke, welche Ergebnisse auf dem Reuters-Korpus erzielt werden können, so dass die hier unternommene Untersuchung in diesem Licht betrachtet werden kann.

5.2.1 Vorverarbeitung

Joachims legt seinen Untersuchungen die Trainings- und Testmenge des ModApte-Splits zugrunde (9603 bzw. 3299 Dokumente respektive). Die Daten wurden von ihm allerdings weiter beschränkt: Es wurden nur Dokumente in das Testverfahren mit

[4] zuletzt unter http://www.daviddlewis.com/resources/testcollections/reuters21578/

aufgenommen, deren Themen mindestens einmal in der Trainings- und Testmenge vorkommen.

Die Dokumente wurden lemmatisiert und alle Stop-Worte entfernt. Es wurden nur solche Worte in die Klassifikation mit aufgenommen, die in mindestens drei Trainingsdokumenten vorkamen[5]. Anschließend wurde für jedes Thema ein Klassifikator trainiert. Zu beachten ist, dass die Klassifikatoren, die bei Joachims Verwendung fanden, grundsätzlich dichotomer Natur waren, d.h., sie konnten für das eine Thema, für das sie trainiert wurden, mit einer bestimmten Fehlerrate feststellen, ob es einem unbekannten (d.h., nicht in der Trainingsmenge enthaltenen) Testdokument zugrundeliegt oder nicht.

Die Featurevektoren wurden mit der *idf*[6] gewichtet, um bessere Ergebnisse zu erzielen.

Im Rahmen dieser Arbeit wurden die Trainings- und Testmenge noch stärker beschränkt: Da der in dieser Arbeit entwickelte Klassifikationsprozess für jedes Dokument genau ein oder kein Thema erkennt, wurden aus der Dokumentenmenge des ModApte-Splits alle Dokumente entfernt, die nicht genau ein annotiertes Thema besaßen.

Weitere Überlegungen betrafen die Ausschaltung möglicher Fehlerquellen: Aufgrund des unüberwachten Lernverfahrens können Fehler sowohl durch den Klassifikator als auch durch Fehler im Clusterprozess zustande kommen. Um den Clusteralgorithmus zu entlasten, wurde C – die Anzahl an Kontexten bzw. Dokumenten, in denen ein Wort mindestens vorkommen muss, um in den Berechnungen berücksichtigt zu werden – auf 20 gesetzt[7]. Weiterhin wurden die Trainings- und Testmenge auf die beiden am häufigsten vorkommenden Themen („`earn`“ mit 2840 Trainingsdokumenten und 1076 Testdokumenten und „`acq`“ mit 1596 Trainingsdokumenten und 695 Testdokumenten) beschränkt. Dass diese Wortmengen trotzdem einen relativ großen Teil des Korpus abdecken, ist auf die heterogene Verteilung der Themen in dem Korpus zurückzuführen: Immerhin gehören ca. 50% aller Dokumente (1771 Dokumente aus der Testmenge und 4436 aus der Trainingsmenge) des ModApte-Splits zu einem der beiden genannten am häufigsten vorkommenden Themen, während viele Themen über weniger als 100 Dokumente insgesamt verfügen.

Analog zu (Joachims, 1998) wurden auch die Entfernung von Funktionswörtern anhand einer Stopliste so wie eine Lemmatisierung durchgeführt. Zur Lemmatisierung wurde wie auch auf dem BITT-Korpus der TreeTagger (Schmid, 1995) verwendet.

Die neue Anzahl der Dimensionen für den LSA-Raum wurde wie in Abschnitt 3.4.3 auf Seite 50 skizziert bestimmt.

5.2.2 Evaluationsergebnisse

Für die beschriebene Teilmenge des ModApte-Splits wurde das Themenerkennungssystem unüberwacht trainiert. Ein zusätzlicher Segmentierungsprozess war im Ge-

[5] Im Rahmen dieser Arbeit wird die Anzahl der Kontexte/Dokumente, in denen ein Dokument mindestens vorkommen muss, mit C bezeichnet. Für diesen Fall ist also C=3. Für eine theoretische Diskussion dieses Wertes siehe Abschnitt 3.7 auf Seite 62.

[6] vgl. Abschnitt 3.31 auf Seite 64

[7] Dieser Wert wurde willkürlich gewählt. Zu beachten ist, dass C relativ zu der Gesamtgröße der Trainingsmenge betrachtet werden muss und somit nicht mit einem C_{rel} von 20 in der BITT-Evaluation gleichgesetzt werden kann.

gensatz zum BITT-Korpus nicht nötig – die zu klassifizierenden Einheiten waren mit den Trainingssegmenten identisch, es handelte sich dabei um die jeweiligen Artikel.

Wie in Abschnitt 5.2.1 beschrieben wurde C auf 20 gesetzt und in Analogie zu (Joachims, 1998) wurden die Häufigkeiten der vorkommenden Worte mit *idf* gewichtet. Das Themenerkennungssystem wurde gezwungen, für jedes Dokument, in dem mindestens ein bekanntes, themenanzeigendes Symbol existierte, ein Thema zu erkennen, indem die Schwelle für die Zurückweisung[8] auf ∞ gesetzt wurde. Weiterhin wurde das Clusterverfahren fest auf zwei Themen eingestellt. Das in Abschnitt 5.1 beschriebene Problem der Zuordnung automatisch erkannter Themen zu manuell annotierten ließ sich aufgrund der überschaubaren Menge an Themen einfach dadurch lösen, dass die beiden automatisch gefundenen Themen je einem der beiden Themen des Korpus manuell zugewiesen wurden.

Tabelle 5.2 zeigt die Akkuratheitswerte des Themenerkennungssystems unter den genannten Parametern.

	Vektor	**Rieger**	**LSA**
Insgesamt	1771	1771	1771
Korrekt	1653	1483	1619
Inkorrekt	99	269	133
Nicht	19	19	19
% korrekt	93.3	83.7	91.4

Tabelle 5.2: Akkuratheit bei der Erkennung von `earn` und `acq`

Die Werte zeigen, dass sowohl das unmodifizierte Vektorraumverfahren als auch der mit LSA modifizierte Vektorraum näherungsweise gleich gute Ergebnisse liefern. Überraschend ist, dass das einfache Vektorraumverfahren sogar besser als der mit LSA modifizierte Raum[9] abschneidet. Ein Erklärung konnte dafür nicht gefunden werden, allerdings hat die Reduktion des Korpus auf zwei häufig repräsentierte Themen und der relativ hohe Wert für C wohl die Störungen drastisch reduziert, die die LSA üblicherweise besser verarbeitet als ein einfaches Vektorraumverfahren.

Das weniger gute Abschneiden des Rieger-Raumes findet sich teilweise auch in späteren Evaluationen[10]. Eine Erklärung hierfür konnte an dieser Stelle nicht gefunden werden.

Die Tabellen 5.3, 5.4 und 5.5 zeigen ein sehr ähnliches Bild wie die Analyse der Akkuratheitswerte.

	Precision	**Recall**	**f**
acq	0.874	0.994	0.930
earn	0.996	0.907	0.949

Tabelle 5.3: Ergebnisse Reuters-Korpus: Vektorraummodell

In (Joachims, 1998) konnten für `acq` gemittelte (*microaveraging*) Werte von 0.853 (C4.5-Entscheidungsbaumklassifikator) über 0.959 (Bayes-Klassifikator) bis hin zu

[8] vgl. Abschnitt 3.6 auf Seite 59

[9] zur Parametrisierung der LSA siehe Abschnitt 3.4.3 auf Seite 50

[10] vgl. Abschnitt 5.3

	Precision	**Recall**	**f**
acq	0.941	0.639	0.761
earn	0.807	0.974	0.883

Tabelle 5.4: Ergebnisse Reuters-Korpus: Rieger

	Precision	**Recall**	**f**
acq	0.846	0.997	0.915
earn	0.998	0.883	0.936

Tabelle 5.5: Ergebnisse Reuters-Korpus: LSA

0.985 (SVM), für `earn` dagegen Werte von 0.959 (Bayes-Klassifikator) bis zu 0.985 (SVM) erzielt werden.

5.2.3 Kritik und Diskussion

Aufgrund der sehr unterschiedlichen Voraussetzungen zu den Ergebnissen in (Joachims, 1998) ist keine echte qualitative Vergleichbarkeit gegeben. Auf der einen Seite werden auf dem gesamten Korpus dichotome Klassifikatoren überwacht trainiert, auf der anderen Seite wird ein nicht-dichotomer Klassifikationsprozess unüberwacht, aber dafür auf einer stark eingeschränkten und somit stark vereinfachten Datenbasis getestet. Bei einer steigenden Anzahl von Themen ist davon auszugehen, dass Fehler des Clusteralgorithmus die Ergebnisse stark zum negativen hin beeinträchtigen würden, so dass die überwacht trainierten Verfahren auf jedem Fall dem hier verwendeten Algorithmus überlegen sind, zumal die von Joachims erzielten Ergebnisse im *break even*-Punkt für `earn` und `acq` leicht bis eindeutig über den in dieser Untersuchung erzielten f-Werten liegen.

Trotzdem zeigen die Ergebnisse der Evaluation deutlich, dass die Klassifikation als solche zufriedenstellende Ergebnisse zu liefern in der Lage ist (f $>$0.9). Im Rahmen nachfolgender Untersuchungen wäre es zweifellos interessant zu explorieren, in wiefern sich dichotome Klassifikatoren für die Themenerkennung im Kontext von multimodaler Mensch-Maschine-Kommunikation einsetzen lassen. Im Rahmen dieser Arbeit wurde jedoch zugunsten einer größeren Bandbreite der Untersuchung darauf verzichtet und der beschriebene Klassifikator verwendet, zumal viele dichotome Klassifikatoren erst ab einer bestimmten Größe der Trainingsmenge gute Ergebnisse zu erzielen in der Lage sind und diese Größe oft im Bereich der Mensch-Maschine-Kommunikation nicht gegeben ist.

Ein Ziel dieser Arbeit ist es, einen Klassifikationsprozess zu finden, der prinzipiell sowohl mit einer sehr geringen Trainingsdatenmenge, als auch mit einem größeren Trainingsdatensatz, wie er in multimodaler HRI entstehen kann, funktioniert. Diese Voruntersuchung so wie die folgende Evaluation auf dem BITT-Korpus sollen letzteres belegen. Die Funktionsfähigkeit mit sehr geringen Trainingsdatenmengen wird dann in den *online*-Experimenten gezeigt.

5.3 Evaluation auf dem BITT-Korpus

Nach den Vorversuchen wurde das Themenerkennungssystem auf dem in Kapitel 4 beschriebenen BITT-Korpus evaluiert (Maas u. a., 2006b) (Maas u. a., 2006a). Die Besonderheit des Korpus besteht wie geschildert in der Möglichkeit der Unterstützung eines Themenerkennungsprozesses durch situierte Information, die einem Roboter während einer situierten, multimodalen Mensch-Maschine-Kommunikation zur Verfügung steht.

Im Rahmen dieser Experimente sollten verschiedene Fragen behandelt werden, deren Beantwortung durch die Entwicklung des Themenerkennungssystems und des Korpus überhaupt erst in den Bereich des Möglichen gerückt worden ist. Kernfragen waren insbesondere:

- Ist eine dynamische *online*-Themenerkennung auf einem natürlichsprachlich kommunizierenden Robotersystem überhaupt möglich?
- Wie gute Ergebnisse lassen sich erwarten?
- Welche Auswirkungen hat die Einbindung multimodaler Information?
- Welche Teilverfahren (semantische Räume, Clustern) sind besonders vielversprechend?
- Welche weiteren Parametrisierungen sind Erfolg versprechend?

Um diese Fragen und weitere zu beantworten, wurden verschiedene Evaluationen angestrengt. Sie werden im Folgenden beschrieben.

5.3.1 Ablauf und Vorverarbeitung

Einen Überblick über den grundsätzlichen Ablauf des Evaluierungsprozesses liefert Abbildung 5.1 auf der nächsten Seite. Die Datengrundlage stellte für alle Vorgänge das BITT-Korpus dar.

Grundsätzlich wurde das Themenerkennungssystem mit ggf. multimodal angereicherten Äußerungen aus dem Korpus trainiert, die zuvor durch einen Segmentierungsprozess in – im Prinzip – thematisch zusammenhängende Abschnitte zusammengefasst wurden. Diese Äußerungen sind identisch mit den von der VAD erkannten[11]. Die Bildung der einzelnen Segmente ist experimentspezifisch und wird weiter unten (Abschnitt 5.3.2 bzw. 5.4.3) beschrieben. Die Segmente wurden zur Bildung eines semantischen Raumes, der anschließend durch einen agglomerativ-hierarchischen Clusterprozess in Themencluster unterteilt wurde, verwendet. Die Themencluster dienten dann der Klassifikation der Testmenge. Sofern nicht anders angegeben, wurde in allen Experimenten ein so genanntes *jack-knive*-Verfahren angewendet, bei dem jeder Monolog des Korpus mit den Themenclustern getrackt wurde, die durch das Training mit allen anderen Monologen entstanden sind. Anschließend erfolgte die Evaluierung der Ergebnisse anhand der manuellen Themenannotation des Korpus.

[11] vgl. Kapitel 4.4.2

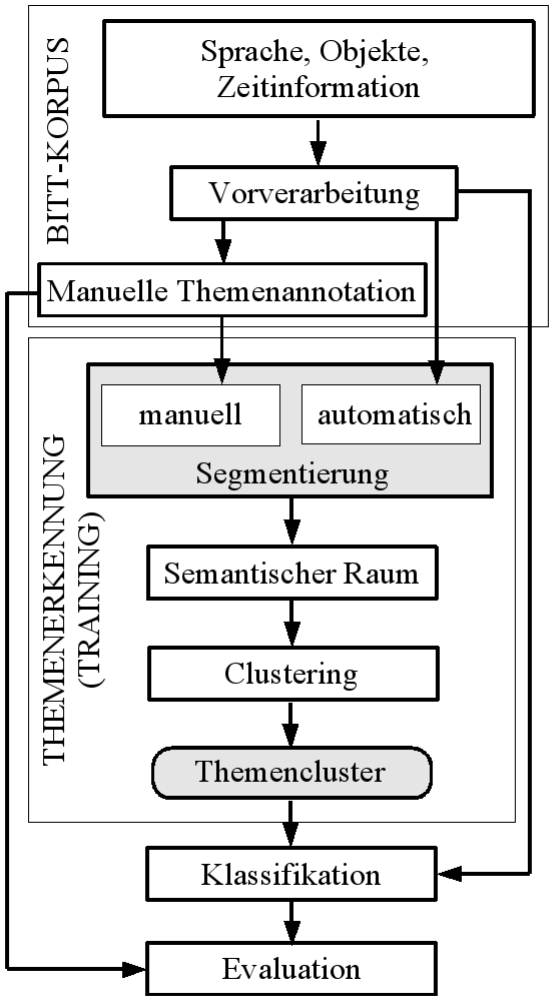

Abbildung 5.1: Datenfluss während der *offline*-Evaluation

5.3.2 Experimente (manuelle Segmentierung)

Eine Experimentreihe diente der Evaluation des Verfahrens unter der Annahme eines perfekten Segmentierungsprozesses. Der Grund für diese Annahme liegt – neben dem Informationsgewinn über die Leistungsfähigkeit des Themenerkennungsverfahrens – in der Schwierigkeit, einen angemessenen Segmentierungsprozess für multimodale HRI zu finden[12]. Weiterhin ist davon auszugehen, dass ein optimaler Segmentierungsprozess auf dem BITT-Korpus aufgrund der Monologstruktur des Korpus andere Ergebnisse erzielen und andere Informationen nutzen würde als in dialogisch-multimodaler HRI.

Um eine möglichst gute Segmentierung der Äußerungen des Korpus zu erzielen, wurde die manuelle Themenannotation als Referenz herangezogen: Direkt aufeinanderfolgende Äußerungen, die dasselbe annotierte Thema besitzen, wurden zu einem Segment zusammengefügt. Äußerungen, die über kein annotiertes Thema verfügten, – wie z.B. die Äußerung „*What next?*“ – wurden dabei aus der Trainingsmenge herausgenommen. Zu beachten ist, dass dies für jede der drei manuellen Annotationen einzeln geschah, so dass Äußerungen oft dreimal, bisweilen seltener, Teil eines Segments der Trainingsmenge waren.

Alternativ hätte aus den manuellen Themenannotationen eine (einzelne) Referenzannotation gewonnen werden können, die als Trainings- und Evaluationsgrundlage gedient hätte. Auf diese Vorgehensweise wurde jedoch verzichtet, da unklar war, nach welchen Kriterien eine solche Referenzannotation hätte gebildet werden sollen.

Parametrisierung

Im Rahmen der Experimente mussten verschiedene Entscheidungen bezüglich der Parametrisierung des Verfahrens getroffen werden. Am wichtigsten war dabei die Konfiguration des Clusteralgorithmus, bei dem auf eine vollständig dynamische Vorgehensweise durch Spezifikation einer maximalen Distanz von Themenclustern verzichtet wurde[13]. Dies geschah aus verschiedenen Gründen: Auch wenn das verwendete Distanzmaß (Korrelation) Hinweise auf die Position eines möglichen Schwellwertes liefert[14] – nämlich im Bereich 0 – ist dennoch bekannt, dass die Anwendung der SVD der LSA die durchschnittliche Distanz im semantischen Raum im Vergleich zum einfachen Vektorraum verringert[15] (Leopold, 2005). Aus diesem Grund konnte nicht für alle Verfahren derselbe Schwellwert verwendet werden; eine individuelle Setzung der Schwellwerte wäre jedoch willkürlich gewesen. Alternativ hätte ein iterativer Prozess zur Findung des idealen Schwellwertes eingesetzt werden können, dies wurde jedoch aufgrund der Länge der Berechnungen so wie der Implausibilität dieses Vorgehens in einer „echten“ Dialogumgebung nicht in Betracht gezogen.

Aus diesem Grund wurde beschlossen, den agglomerativen Clusteralgorithmus automatisch bei einer Schwelle von 10 gefundenen Clustern terminieren zu lassen. Der Wert von 10 ergab sich aus den manuellen Annotationen, da er leicht höher ist als

[12] vgl. Abschnitt 5.4.3

[13] Wie beschrieben kann der agglomerative Clusteralgorithmus dann terminieren, wenn die zu vereinigenden Cluster weiter als eine festgelegte Distanz voneinander entfernt sind.

[14] vgl. dazu Abschnitt 5.4.2

[15] Zur Erinnerung: Ein hoher Wert für die Korrelation entspricht einer niedrigen Distanz im semantischen Raum.

die maximale Anzahl in den Monologen annotierter Themen. Trotzdem konnten weniger als 10 Themen automatisch in den Monologen erkannt werden, wenn keine zu dem jeweiligen Cluster passenden Äußerungen gefunden wurden. Dies war insbesondere im Fall von „Splitterclustern“ der Fall, bei denen ein oder zwei Worte seltener Verwendung zu einem Thema zusammengefasst wurden. Auf die Möglichkeiten real dynamischen Clusterns wird kurz in Abschnitt 5.4.2 eingegangen, so wie in der Darstellung der *online*-Experimente in Abschnitt 6.

Die Wortvektoren der semantischen Räume wurden mit *idf* bzw. Entropie gewichtet. Wie beschrieben wurde zum Training und Testen ein *jack-knife*-Verfahren angewandt, bei dem jeder Monolog mit den 28 anderen Monologen als Trainingsdaten getrackt wurde.

Aufgrund der oben beschriebenen Segmentierung anhand der manuellen Themenannotationen musste eine besondere Behandlung für C – die Menge an Kontexten, in denen ein Symbol mindestens vorkommen muss, um in die Berechnung mit aufgenommen zu werden – vorgenommen werden. Aufgrund der Segmentierung wurden bestimmte Äußerungen nicht mit in die Trainingsmenge aufgenommen, dabei handelte es sich um solche, die keine thementragenden Symbole enthielten, oder um solche, die von dem jeweiligen Annotator als nicht thementragend gekennzeichnet worden waren. Außerdem wurden die Annotationen aller drei Annotatoren berücksichtigt, was zu einer effektiven Verdreifachung des Korpusgröße führte. Aus diesen Gründen konnte C nicht wie gehabt berechnet werden. Zwar existierte der Parameter C nach wie vor für den Algorithmus, aber dieser konnte nicht in Bezug auf andere Evaluationen als vergleichbar gewertet werden. Aus diesem Grund wurde der Vergleichswert C_{rel} eingeführt. Dieser Wert soll ermöglichen, über ein von korpusspezifischen Umständen weitgehend freies Maß zur Bestimmung des Anteils gelöschter, selten vorkommender Symbole zu verfügen. Ein konkretes C berechnet sich aus einem vorgegebenen C_{rel} wie folgt:

$$C = C_{rel} * (n - 1) + 1 \tag{5.5}$$

n ist dabei die Anzahl der in Betracht gezogenen Annotationen, in diesem Fall 3. Auf diese Weise können verschiedene Werte für C/C_{rel} miteinander verglichen werden.

Wie in Abschnitt 3.6 auf Seite 60 beschrieben wurde die *history* aktiviert; der Verfall thematischer Information wurde durch Multiplikation der vorgehenden Themenstärken mit 0.3 bewirkt. In den Fällen, in denen sich für alle möglichen Themen eine negative Summe der Korrelationswerte ergab, wurde das letzte ermittelte Thema (sofern gegeben) wie beschrieben als aktuelles Thema angenommen.

Ein verfahrensspezifischer Parameter ist die Wahl der Dimensionalität für die LSA. In allen Experimenten, in denen dieses Verfahren verwendet wurde, wurde die in Abschnitt 3.4.3 auf Seite 50 skizzierte Vorgehensweise zur dynamischen Bestimmung der Dimensionalität angewandt.

Evaluationsprozess

Die Evaluation erfolgte – in Analogie zu der Evaluation des Reuters-Korpus – anhand zwei verschiedener Maße. Zum einen wurde wiederum eine *precision/recall*-basierte Evaluation mit f als gemitteltem Wert unternommen. Dabei stellte sich wiederum das Problem, dass *precision* und *recall* nur für ein Thema ermittelt werden können, aber ein einzelner Wert für die Leistungsfähigkeit des Verfahrens ermittelt werden

sollte. Aus diesem Grund sind die in den Tabellen und Grafiken angegebenen f-Werte arithmetische Mittelwerte aus den jeweiligen Einzelwerten, wobei sämtliche f-Werte aller Themen und Annotationen herangezogen wurden. Die f-Werte sind somit *nicht* nach der jeweiligen Vorkommenshäufigkeit des jeweiligen Themas gewichtet.

Um jedem manuell annotierten Thema ein passendes automatisch gebildetes Thema zuzuordnen, wurde – wie in Abschnitt 5.1 beschrieben – jedem manuell annotierten Thema das jeweilige am häufigsten in den Trackingergebnissen übereinstimmende automatisch generierte Thema zugeordnet.

Um einen ungefähren Eindruck der relativen Häufigkeit korrekter Trackingergebnisse zu gewinnen, wurde eine zweite Evaluation durchgeführt. Wie weiter oben (Abschnitt 5.1) dargestellt, ist das übliche Maß zur Angabe des Anteils korrekter Ergebnisse eines Themenerkennungssystems Akkuratheit. Dieses Maß kann jedoch in bestimmten Fällen zu fehlerhaften Ergebnissen führen. Unter den gegebenen Umständen wäre dies insbesondere der Fall gewesen, wenn während des Klassifikationsprozesses nur ein Thema gefunden worden wäre, welches einen Großteil des Korpus abgedeckt hätte. Während verschiedener Testläufe des Klassifikationsprozesses trat dieser Fall häufig dann ein, wenn der Clusteralgorithmus aufgrund einer zu schlechten Datenlage keine echten Cluster finden konnte. Häufig wurden in diesem Fall viele Ein-Wort-Cluster gefunden, so wie ein oder zwei weitere, die die restlichen Symbole abdeckten.

Aus diesem Grund wurde ein alternatives Evaluationsmaß – **invertierte Akkuratheit** – entwickelt, welches Nutzen aus der verhältnismäßig gleichmäßigen Verteilung der manuell annotierten Themen im Korpus zieht: Im Prinzip entspricht eine Evaluation nach invertierter Akkuratheit einer Evaluation nach Akkuratheit, jedoch wurde gemessen, wie gut die manuell annotierten Themen die automatisch gefundenen abdecken. (Bei einer Evaluation nach Akkuratheit wäre gemessen worden, wie gut die automatisch gefundenen Themen die manuell annotierten abdecken.) Dabei wurden jedoch zwei Modifikationen vorgenommen: Fälle, in denen kein manuell annotiertes Thema vorlag, wurden ignoriert und Fälle, in denen ein manuell annotiertes Thema aber kein automatisch gefundenes vorlag, wurden als „falsch“ gekennzeichnet.

Auf diese Weise ließ sich das beschriebene Problem umgehen, aber trotzdem ein Maß für den relativen Anteil der korrekt klassifizierten Äußerungen anwenden.

Die folgenden Abschnitte beschreiben und diskutieren die Ergebnisse der einzelnen Auswertungen mit den jeweiligen Evaluationsmaßen. Im Vorfeld jedoch wurden anhand verschiedener Experimente die *baseline* und das *upper limit* – die obere und untere Grenze – ermittelt, die sich mit den vorgestellten Verfahren auf den Daten erreichen liessen.

Baseline- und upper-limit-Bestimmung

Auch im Fall eines maximal schlecht arbeitenden Klassifikators erzielen die vorgestellten Evaluationsmaße aufgrund inhärenter Kriterien aber auch aufgrund von *history*-verursachten Glättungsprozessen nicht 0%. Aufgrund dieses Umstandes ist es wichtig zu erfahren, welche Ergebnisqualität sich ohne Zutun des Klassifikators einstellt.

Genauso ist aufgrund des Umstandes, dass die manuellen Annotationen nicht völlig übereinstimmen, unter der gegebenen Evaluationsmethode kein maximales Ergebnis von 100% möglich, da gegen alle drei manuellen Annotationen evaluiert und gemittelt wird. Wie oben beschrieben (Abschnitt 5.3.2) hätte diesem Umstand Abhilfe

geschaffen werden können, indem aus den drei manuellen Annotationen eine Referenzannotation gewonnen worden wäre. Die Kriterien zur Erstellung einer solchen Annotation wären jedoch willkürlich gewesen – z.B. ist unklar, ob eine solche Referenzannotation aus der Menge der minimalen Übereinstimmungen oder aus einer größeren Menge von Daten gebildet werden müsste. Aus diesem Grund wurde darauf verzichtet und stattdessen ein *inter rater agreement*, ein Maß für die Übereinstimmung der Annotationen ermittelt. Dieses stellt im vorliegenden Fall auch gleichzeitig eine – wenn auch weiche – obere Grenze für das maximal von dem automatischen Klassifikationsprozess zu erzielenden Ergebnis dar.

Die folgenden Abschnitte schildern die jeweiligen Werte und Vorgehensweisen bei der Ermittlung derselben.

Baseline Es existieren mehrere Möglichkeiten, die untere Grenze für die Evaluationsergebnisse zu bestimmen. Wichtig war vor allem die Erfassung von Effekten wie Glättung durch die Anwendung der *history*. Ein Verfahren, bei dem die Äußerungen zufallsbasiert den möglichen Themen zugeordnet werden würden, wäre somit nicht angemessen gewesen, da das gewählte Evaluationsverfahren homogene Klassifikationen besser bewertet.

Im Vorfeld stellte sich die Frage, welche weiteren Faktoren neben der Verwendung der *history* die untere Grenze zu beeinflussen in der Lage wären. Da davon auszugehen war, dass sich Themen in der manuellen Annotation grundsätzlich über mehrere Zeilen erstreckten, kamen Parameter in Frage, die zu weniger Varianz in der automatischen Erkennung der Themen führen. Beispiele dafür sind C bzw. C_{rel}, da diese die Anzahl an möglichen Stellen begrenzen, an denen ein Themenwechsel stattfinden kann. Aber auch andere Gründe, aus denen sich die Anzahl oder Gruppe der zur Themenerkennung herangezogenen Symbole ändern konnte, sind kritisch. So zum Beispiel die Stopliste – die aber zu keinem Zeitpunkt modifiziert wurde – und vor allem die Art der zum Training und Tracking verwendeten Symbole. Aus diesem Grund mussten alle diese Faktoren entsprechend den Experimenten variiert und zur Berechnung einzelner Messwerte für den jeweiligen Parametersatz herangezogen werden.

Auf zufälligen Daten sollten Faktoren wie die Art des verwendeten semantischen Raumes zwar theoretisch in kleinem Maße das Ergebnis des Clusteralgorithmus modifizieren und somit einen Einfluss auf das Ergebnis haben, dieser Aspekt wurde jedoch aufgrund des zu erwartend geringen Einflusses nicht in die Berechnungen mit aufgenommen.

Zur Berechnung der unteren Grenze wurde wie folgt vorgegangen: Ein semantischer Raum wurde mit allen Monologen des Korpus trainiert und geclustert. Aufgrund der im Folgenden beschriebenen Verwürfelung erschien die Verwendung eines *jack-knife*-Verfahrens unnötig. Als Typ des semantischen Raumes wurde die LSA verwendet. In den resultierenden Themenclustern wurden im Anschluss alle Symbole durch ein zufällig bestimmtes anderes Symbol ersetzt. Dabei wurde darauf geachtet, dass ein Symbol stets durch dasselbe Symbol substituiert wurde, so wie dass kein Symbol zweimal als Substituent herangezogen wurde. Anschließend wurden die Themencluster dazu verwendet, die Monologe zu tracken und das Ergebnis wurde wie in den eigentlichen Evaluationen mit den manuellen Annotationen verglichen.

Der Prozess wurde für jeden Parametersatz – Art der Referenzen und Wert für C_{rel} – wiederholt. Die Ergebnisse für die f-Wert-basierte Evaluation so wie für die

Evaluation mit invertierter Akkuratheit finden sich in den Tabellen 5.6, 5.7 und 5.8. Eine genauere Beschreibung der zugrundeliegenden Symbolmengen (wortbasiert, objektreferenzbasiert, gemischt) findet sich in den jeweiligen, folgenden Abschnitten, in denen die Evaluationsergebnisse diskutiert werden.

Grundsätzlich ist erkennbar, dass die arithmetisch gemittelten *recall*-Werte wesentlich höher sind als die ermittelten Werte für *precision*. Dies ist dadurch erklärbar, dass mehrere manuell annotierte Themen von einzelnen automatisch erkannten zusammengefasst wurden. Der daraus resultierende geringe Anstieg von f stützt diese These, da er darauf hindeutet, dass die *recall*-Mittelwerte insbesondere durch Fälle angehoben wurden, in denen sehr hohe *recall*-Werte mit geringen *precision*-Werten gekoppelt waren. Genau dies wäre bei einer Zusammenfassung mehrerer manueller Themen in einem automatischen Thema der Fall.

Einem Quervergleich der Tabellen ist zu entnehmen, dass wie erwartet die unteren Grenzen für den rein objektbasierten Ansatz höher liegen als für die anderen beiden Ansätze, insbesondere für den rein wortbasierten Ansatz. Dies ist erklärbar durch die geringere Anzahl von Symbolen, die dem gemischten und dem rein objektreferenzbasierten Ansatz zugrunde liegen. Derselbe Effekt stellt sich erwartungsgemäß durch die Erhöhung von C_{rel} ein. Prinzipiell lassen die Ergebnisse bei sehr hohen Werten für C_{rel} darauf schließen, dass in diesem Bereich Evaluationen sehr wenig aussagekräftig sind. Stabile Evaluationen sind aber mindestens im Bereich von einem C_{rel} von 2 bis 10 möglich.

Upper limit Die Ermittlung der oberen Grenze bzw. des *inter rater agreements* wurde einfach durch die wechselseitige Anwendung des jeweiligen Evaluationsmaßes auf die manuellen Annotationen ausgeführt. So wurde sechsmal – für jedes der drei Paare manueller Annotationen und jede Richtung – angenommen, die eine Annotation sei die Referenzannotation und die andere das Ergebnis eines zu evaluierenden Klassifikationsprozesses. Im Fall der inversen Akkuratheits-Evaluation wurde für die drei Annotationen ein arithmetisch gemittelter Wert von 0.95 erreicht, was einer relativ hohen Übereinstimmung entspricht. Im Fall der wechselseitigen Evaluation mit Hilfe eines ungewichteten f-Wertes wurde zunächst 0.79 als obere Grenze ermittelt. Im Gegensatz zu dem automatischen Klassifikationsvorgang konnte das erzielte Ergebnis jedoch aufgrund des Fehlens einer *history*-basierten Glättung zustande kommen, also in Fällen, in denen in einer Annotation eine geringere Menge an Äußerungen annotiert wurde als in einer anderen. Da das automatische Themenerkennungsverfahren in nahezu jedem Fall ein Thema erkannte – im Zweifelsfall das letzte echt erkannte – schien der ermittelte Wert nicht exakt der gesuchten oberen Grenze zu entsprechen.

Aus diesem Grund wurde ein weiteres mal evaluiert, in diesem Fall wurden jedoch – sofern möglich – alle nicht annotierten Äußerungen in der „automatischen“ Annotation mit dem vorangegangenen Thema versehen. Ergebnis war allerdings nur ein minimaler Anstieg des resultierenden gemittelten f auf 0.80.

Nach Darlegung der oberen und unteren Grenzen für die Evaluation möchte ich im Folgenden auf die erzielten Ergebnisse eingehen.

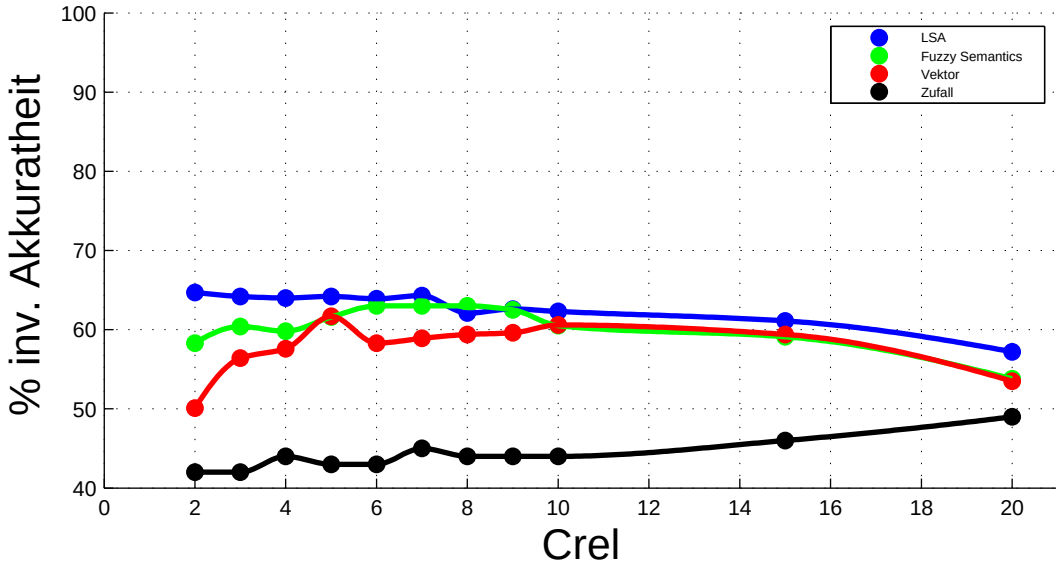

Abbildung 5.2: Grafische Darstellung der Evaluationsergebnisse, rein lemmabasierter Ansatz, invertierte Akkuratheit

Wortbasierter Ansatz

In ersten Experimenten wurde ein unimodaler Ansatz verfolgt, bei dem als Datenbasis ausschließlich die in den Monologen geäußerten, lemmatisierten Worte dienten. Auf diese Weise konnte im Vergleich zu den multimodalen Ansätzen ermittelt werden, wie sehr sich in situierter Kommunikation multimodale Indizien verbessernd auswirken können. Die Tabellen 5.9, 5.10 und 5.11 (grafische Darstellung in den Abbildungen 5.2 und 5.3) zeigen die jeweiligen Ergebnisse, wobei jeder der drei semantischen Räume (einfacher Vektorraum, Rieger, LSA) getestet wurde.

Es ist leicht ersichtlich, dass die Verfahren bezüglich aller Evaluationsmaße zwar eindeutig oberhalb der ermittelten unteren Grenzen gelagerte Resultate erzielen, jedoch sind sie in keiner Weise als befriedigend zu bewerten. Gründe dafür mögen in den themenübergreifend verwendeten Objektbezeichnungen liegen – so z.B. die Benennung zweier thematisch verschiedener Stühle als „Stuhl“ – so dass der Clusteralgorithmus schlechte Ergebnisse erzielt. Möglich ist auch die Verwendung von halbindividuellen Bezeichnungen durch die Versuchspersonen. Ein Beispiel dafür wäre die Bezeichnung der Küchenregion als „kitchenette“ oder als „kitchen“. Aufgrund der geringen Datengrundlage lassen sich diese Begriffe nicht unbedingt sinnvoll maschinell Themen zuordnen.

Zwischen den einzelnen Räumen lässt sich anhand dieser Ergebnisse kein einschlägiger Unterschied bezüglich der Leistungsfähigkeit feststellen, obgleich die LSA im Durchschnitt die besten Resultate sowohl bezüglich f als auch der invertierten Akkuratheit liefert.

Gemischter Ansatz

Die Tabellen 5.12, 5.13 und 5.14 (grafische Darstellung in den Abbildungen 5.4 und 5.5) zeigen die Ergebnisse der ersten multimodalen Evaluationen. In diesen Experi-

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.28	0.27	0.27	0.26	0.25	0.26	0.26	0.27	0.28	0.27	0.32
recall	0.40	0.41	0.44	0.46	0.47	0.47	0.48	0.47	0.49	0.54	0.55
f	0.28	0.28	0.29	0.29	0.28	0.29	0.29	0.29	0.31	0.30	0.34
% inv. Akkur.	42	42	44	43	43	45	44	44	44	46	49

Tabelle 5.6: Baseline – rein wortbasierter Ansatz

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.31	0.26	0.28	0.27	0.26	0.26	0.27	0.28	0.28	0.36	0.40
recall	0.41	0.47	0.46	0.49	0.48	0.49	0.51	0.51	0.51	0.57	0.65
f	0.31	0.29	0.30	0.30	0.30	0.29	0.31	0.31	0.31	0.38	0.43
% inv. Akkur.	42	44	43	45	44	45	45	47	49	48	52

Tabelle 5.7: Baseline – gemischter Ansatz

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.31	0.29	0.31	0.31	0.30	0.37	0.32	0.35	0.28	0.35	0.41
recall	0.50	0.50	0.53	0.55	0.57	0.56	0.57	0.58	0.58	0.63	0.74
f	0.32	0.32	0.33	0.34	0.33	0.38	0.36	0.37	0.33	0.39	0.45
% inv. Akkur.	47	46	46	49	48	51	49	54	48	55	60

Tabelle 5.8: Baseline – rein objektreferenzbasierter Ansatz

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.34	0.41	0.45	0.51	0.46	0.46	0.45	0.46	0.48	0.44	0.36
recall	0.72	0.71	0.73	0.70	0.74	0.74	0.72	0.71	0.70	0.70	0.70
f	0.40	0.46	0.49	0.53	0.50	0.50	0.49	0.50	0.50	0.48	0.41
% inv. Akkur.	51	56	58	62	58	59	59	60	61	59	54

Tabelle 5.9: Einfacher Vektorraum, rein wortbasiert

$\boldsymbol{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.46	0.50	0.47	0.49	0.52	0.52	0.50	0.50	0.49	0.46	0.36
recall	0.61	0.63	0.63	0.62	0.65	0.68	0.67	0.67	0.67	0.66	0.64
f	0.46	0.50	0.47	0.48	0.52	0.53	0.52	0.52	0.51	0.49	0.41
% inv. Akkur.	58	60	60	62	63	63	63	62	61	59	54

Tabelle 5.10: Rieger, rein wortbasiert

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.54	0.55	0.54	0.56	0.53	0.54	0.52	0.52	0.51	0.50	0.42
recall	0.63	0.64	0.65	0.64	0.66	0.66	0.65	0.65	0.65	0.65	0.63
f	0.52	0.53	0.53	0.54	0.53	0.54	0.52	0.52	0.51	0.50	0.44
% inv. Akkur.	65	64	64	64	64	64	63	63	62	61	57

Tabelle 5.11: LSA, rein wortbasiert

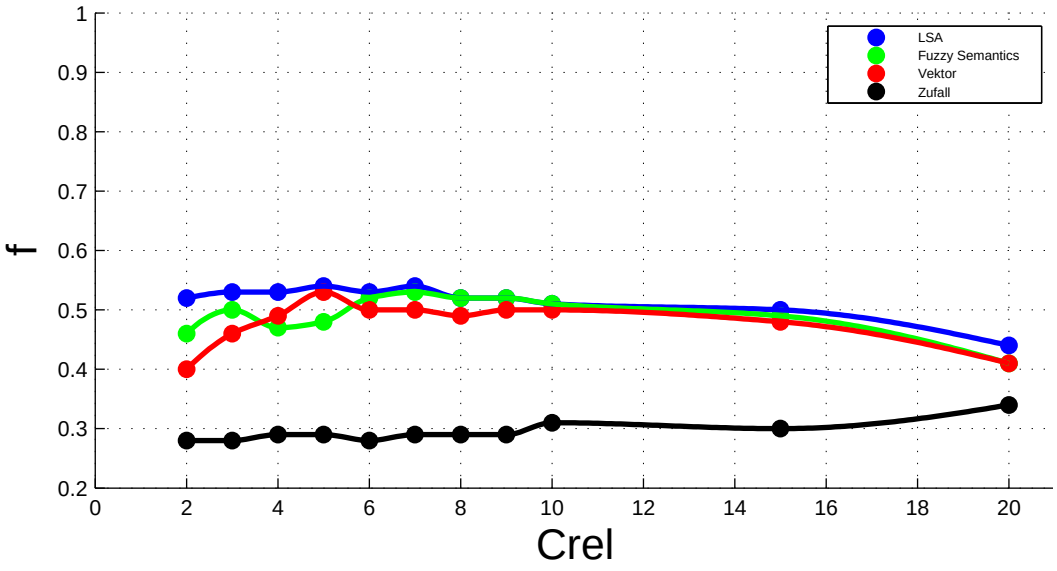

Abbildung 5.3: Grafische Darstellung der Evaluationsergebnisse, rein lemmabasierter Ansatz, gemitteltes f

menten wurden Objekt- und Gruppen-IDs mit in die Trainingsmenge aufgenommen. Grundsätzlich wurden dabei die verbalen Objektreferenzen (sofern vorhanden) durch die IDs ersetzt; so wurde z.B. der Satz:

Beispiel 5.1 *look at the little raven*

durch die Symbolkette

Beispiel 5.2 *look at raven_01*

substituiert. Prinzipiell hätte die Objekt-ID auch einfach hinzugefügt werden können; auf diese Vorgehensweise wurde aber verzichtet, da von einer Verschlechterung des Clusterergebnisses aufgrund der eine hohe Polytextie aufweisenden Begriffe ausgegangen wurde.

Zu bedenken ist, dass auf diese Weise zumindest scheinbar Informationen über Zugehörigkeit von Objekttypen zu Themen („Schwämme“ gehören in die Küche und in das Badezimmer) verloren gehen – die Phrase „*the little raven*“ enthält zumindest in der englischen Sprache aufgrund des Substantivs begrenzt Information darüber, welcher Objekttyp beteiligt ist, nämlich „*raven*“. Diese scheinbar gelöschten Informationen können jedoch in einem späteren Schritt wieder aus den Daten gewonnen werden[16].

Gruppen- und Objekt-IDs wurden gleich behandelt, spontan gebildete Gruppen wurden durch ihre Mitglieder repräsentiert.

Für die vorliegenden Resultate wurden die zum Tracking herangezogenen Verbindungsstärken der Worte (im Gegensatz zu denen der Referenzen) mit einem Strafmultiplikator von 0.5 belegt.

[16]So kann z.B. aus dem Wissen, dass `schwamm_01` und `schwamm_02` Schwämme sind und sie zu den Themen „Küche“ respektive „Bad“ gehören, wieder auf die Zugehörigkeit des Objekttyps „Schwamm“ zu den jeweiligen Themen geschlossen werden.

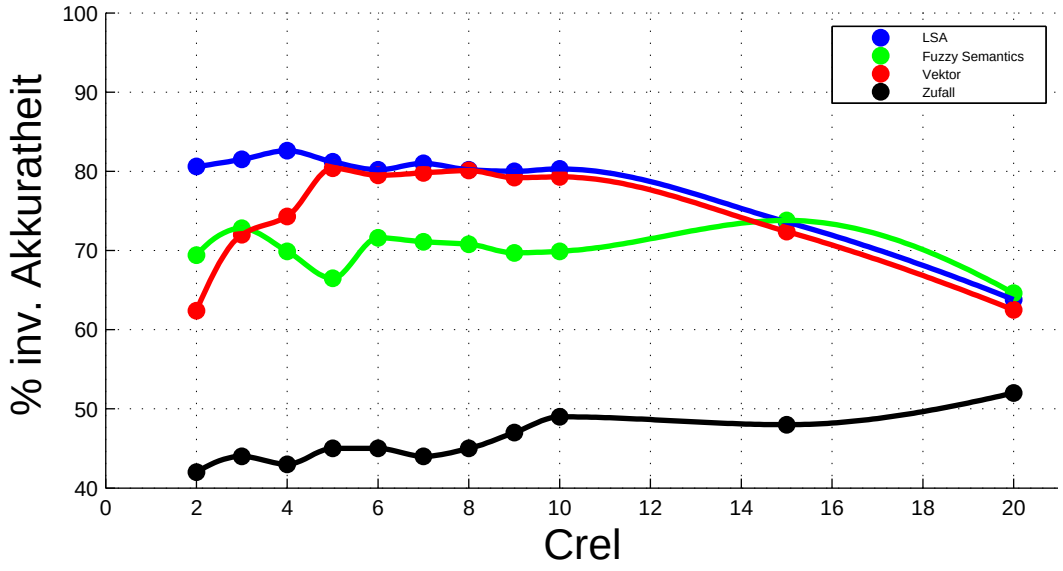

Abbildung 5.4: Grafische Darstellung der Evaluationsergebnisse, gemischter (Lemmata + Objektreferenzen) Ansatz, invertierte Akkuratheit

Die Versuchsreihen wurden abgesehen von den genannten Punkten unter denselben Parametern wie im vorigen Abschnitt aufgenommen.

Im Vergleich zu dem rein wortbasierten Ansatz lassen sich schon gravierende Verbesserungen erkennen. Interessant ist das schlechtere Abschneiden des Rieger-Raums, da es – auch im Vorgriff auf die noch darzustellenden Evaluationen – nur mit der Mischung von Symbolen aus zwei Modalitäten einhergeht. Ggf. deutet dies darauf hin, dass der semantische Raum nach Rieger nur für eine einheitliche Verteilung von Symbolen, aber nicht wie im vorliegenden Fall für die Mischung von zwei Verteilungen geeignet ist. Leider fehlen aber an dieser Stelle weitere Erkenntnisse, so dass es sich hierbei um reine Spekulation handelt.

Sehr auffällig ist die Tatsache, dass der einfache Vektorraum zwar im Bereich niedriger Werte für C_{rel} deutlich schlechtere Ergebnisse liefert als die LSA, sich dies aber im Bereich ab einem C_{rel} von 5 gibt. Dieses Ergebnis lässt schließen, dass die von der LSA beseitigten störenden Einflüsse – denen die LSA aufgrund des beschriebenen Linearisierungsprozesses besser gewachsen ist als ein einfacher Vektorraum – hauptsächlich in Form von selten vorkommenden Symbolen vorliegen. Auf diese Weise kann durch Löschung dieser Symbole aus der Trainingsmenge das einfache Vektorraumverfahren vergleichbar gute Ergebnisse wie die LSA liefern.

Aufgrund des seltenen Vorkommens der gelöschten Symbole in der Trainingsmenge ist eine starke Beeinflussung des Trackings der Testmenge durch diese Symbole mit hoher Wahrscheinlichkeit ausgeschlossen. Viel wahrscheinlicher ist wiederum eine negative Beeinflussung des Clusterprozesses. Ggf. kann das Verfahren durch einen robusteren Clusteralgorithmus oder eine stärkere Gewichtung der Symbole nach Häufigkeit positiv beeinflusst werden.

Die in diesen Versuchsreihen erzielten Ergebnisse sind zwar noch nicht optimal, können jedoch schon eher als Grundlage für an die Themenerkennung anschließende Prozesse genutzt werden als die rein wortbasierten Resultate.

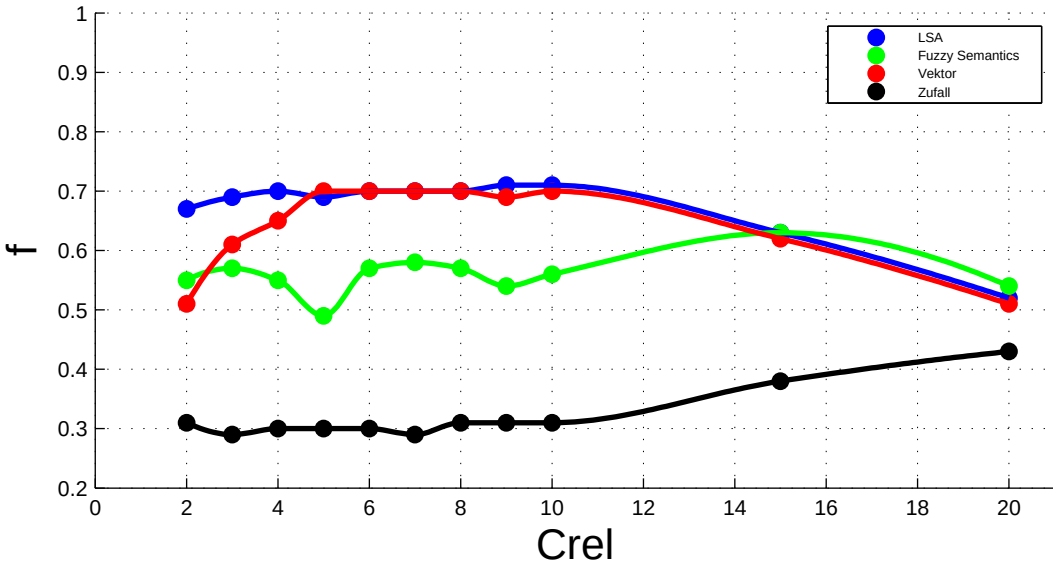

Abbildung 5.5: Grafische Darstellung der Evaluationsergebnisse, gemischter (Lemmata + Objektreferenzen) Ansatz, gemitteltes f

Referenzbasierter Ansatz

In einer weiteren Versuchsreihe wurden ausschließlich Objektreferenzen (bzw. Gruppenreferenzen) in die Trainingsmenge mit aufgenommen. Der in der gemischten Versuchsreihe angedeutete Umstand, dass eine starke Bindung von Objekten und Themen in situierter Kommunikation besteht, sollte auf diese Weise noch einmal überprüft werden.

Die Ergebnisse in den Tabellen 5.15, 5.16 und 5.17 (grafische Darstellungen in den Abbildungen 5.6 und 5.7) bergen zwei Überraschungen: Zum einen lassen sich mit allen Verfahren Resultate erzielen, die sehr nahe an bzw. sogar auf den ermittelte Obergrenzen von 0.80 f bzw. 0.95 für invertierte Akkuratheit liegen. Dies kann als eindeutiges Indiz dafür gewertet werden, dass situierte Information notwendig für Themenerkennung in situiert-multimodaler Mensch-Roboter-Kommunikation ist. Zum anderen sind die im gemischten Verfahren anfänglich aufgetretenen Einbußen des einfachen Vektorraums gegenüber der LSA nicht mehr zu erkennen. Offenbar ist die Datengrundlage aufgrund der Kombination der Verwendung von Objektreferenzen und der manuellen Segmentierung so eindeutig, dass durch die LSA keine Störfaktoren beseitigt werden. Weiterhin sind die Ergebnisse des Fuzzy Semantics-Verfahrens wieder auf dem Niveau der anderen beiden. Dies ist möglicherweise der Fall, weil wiederum nur eine einzelne Modalität verarbeitet wird.

Auswirkungen von C_{rel}

Nachdem in den vorangegangenen Abschnitten die Auswirkungen von C_{rel} auf die Qualität der Evaluation eingegangen wurde, war es interessant zu erfahren, wie sich die semantischen Räume im Detail bei Veränderungen dieses Wertes verhalten. Dabei waren vor allem die Größe der jeweiligen semantischen Räume in Form von Symbo-

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.48	0.59	0.64	0.71	0.71	0.71	0.71	0.7	0.71	0.6	0.48
recall	0.77	0.78	0.79	0.78	0.79	0.79	0.79	0.78	0.79	0.76	0.75
f	0.51	0.61	0.65	0.7	0.7	0.7	0.7	0.69	0.7	0.62	0.51
% inv. Akkur.	81	82	83	81	80	81	80	80	80	74	64

Tabelle 5.12: Einfacher Vektorraum, gemischter Ansatz

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.51	0.53	0.5	0.42	0.53	0.53	0.52	0.49	0.51	0.63	0.51
recall	0.77	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.75	0.74
f	0.55	0.57	0.55	0.49	0.57	0.58	0.57	0.54	0.56	0.63	0.54
% inv. Akkur.	69	73	70	67	72	71	71	70	70	74	65

Tabelle 5.13: Rieger, gemischter Ansatz

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.68	0.69	0.72	0.7	0.7	0.72	0.72	0.72	0.72	0.63	0.49
recall	0.76	0.78	0.78	0.77	0.78	0.77	0.79	0.78	0.78	0.75	0.74
f	0.67	0.69	0.7	0.69	0.69	0.7	0.7	0.7	0.71	0.63	0.52
% inv. Akkur.	81	82	83	81	80	81	80	80	80	74	64

Tabelle 5.14: LSA, gemischter Ansatz

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.79	0.79	0.75	0.78	0.73	0.73	0.71	0.71	0.71	0.64	0.55
recall	0.89	0.89	0.88	0.85	0.88	0.88	0.88	0.88	0.88	0.84	0.8
f	0.8	0.8	0.76	0.77	0.75	0.75	0.74	0.74	0.74	0.67	0.59
% inv. Akkur.	90	90	89	90	88	88	87	86	86	83	72

Tabelle 5.15: Einfacher Vektorraum, nur Objektreferenzen

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.68	0.76	0.76	0.78	0.77	0.77	0.77	0.76	0.77	0.63	0.54
recall	0.9	0.9	0.89	0.88	0.88	0.88	0.88	0.88	0.87	0.84	0.8
f	0.72	0.78	0.78	0.79	0.78	0.78	0.78	0.77	0.78	0.67	0.57
% inv. Akkur.	86	89	89	90	89	89	88	88	88	82	72

Tabelle 5.16: Rieger, nur Objektreferenzen

$\mathbf{C}_{rel}$	2	3	4	5	6	7	8	9	10	15	20
precision	0.77	0.79	0.77	0.77	0.76	0.75	0.74	0.74	0.74	0.63	0.55
recall	0.86	0.87	0.86	0.86	0.86	0.86	0.87	0.86	0.87	0.84	0.8
f	0.78	0.79	0.77	0.77	0.76	0.76	0.75	0.75	0.75	0.67	0.59
% inv. Akkur.	90	90	89	89	88	88	88	87	87	83	72

Tabelle 5.17: LSA, nur Objektreferenzen

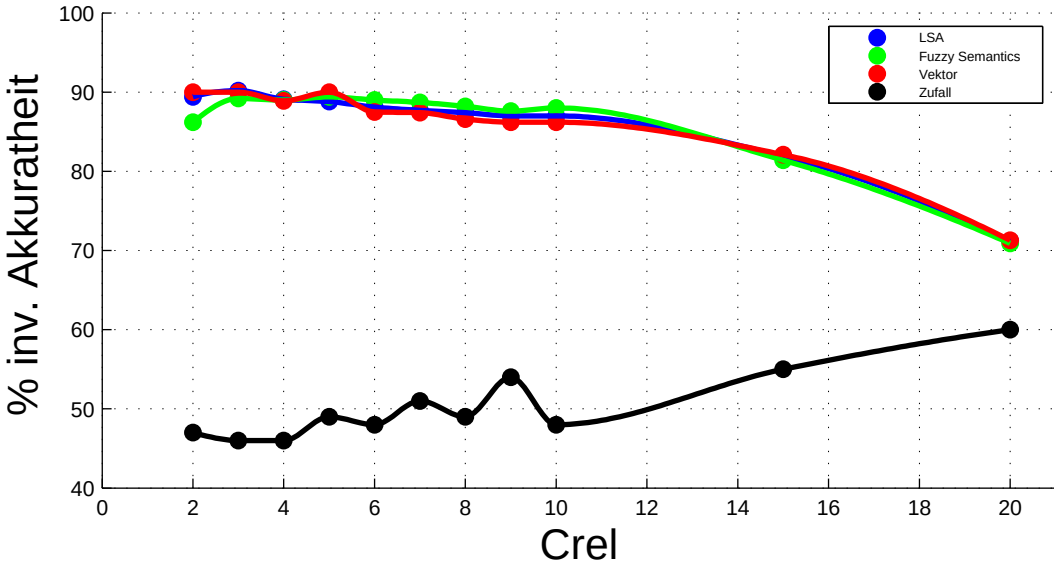

Abbildung 5.6: Grafische Darstellung der Evaluationsergebnisse, rein objektreferenzbasierter Ansatz, invertierte Akkuratheit

len von Interesse, da diese der Anzahl an Themen zugeordneten Symbolen entspricht, die in weiteren Anwendungen nutzbar sind. Ein weiterer interessanter Aspekt war die relative Anzahl von Äußerungen, die nur aufgrund der *history* einem Thema zugeordnet wurden, da auf diese Weise ein Einblick in die Relevanz der *history* gewonnen werden kann.

Symbolanzahl nach C_{rel} Eine wichtige Frage ist die Stärke der Auswirkung von C bzw. C_{rel} auf die Größe der Menge der im semantischen Raum vertretenen Symbole. Abgesehen von C_{rel} wirken sich dabei noch zwei weitere Faktoren aus: Zum einen die Größe der gebildeten Segmente und zum anderen die zugrundeliegende Symbolmenge. Letztere ergibt sich aus der Datenquelle und der Art (Worte, Lemmata, ObjektIDs) der verwendeten Symbole. Die Auswirkungen der Symbolmenge entstehen durch die unterschiedlichen, den Symboltypen zugrundeliegenden Verteilungsprinzipien - so kann z.B. die Substitution von Nominalphrasen durch Objekt-IDs einzelne, häufig vorkommende Symbole („Tasse“) durch mehrere, selten vorkommende (`cup_01, cup_02 ...`) ersetzen, die dann bei niedrigeren Werten von C_{rel} gelöscht werden.

Tabelle 5.18 gewährt einen Eindruck, wie viele Symbole als Grundlage für Klassifikation und Weiterverarbeitung bei den jeweiligen Werten für C bzw. C_{rel} übrig bleiben.

C_{rel}	2	3	4	5	6	7	8	9	10	15	20
wortbasiert	587	397	305	245	212	181	158	141	124	80	57
gemischt	633	427	324	256	216	185	160	142	126	72	36
Objektreferenzen	195	162	138	116	104	94	80	73	67	40	17

Tabelle 5.18: Symbolanzahl in Abhängigkeit von C_{rel}

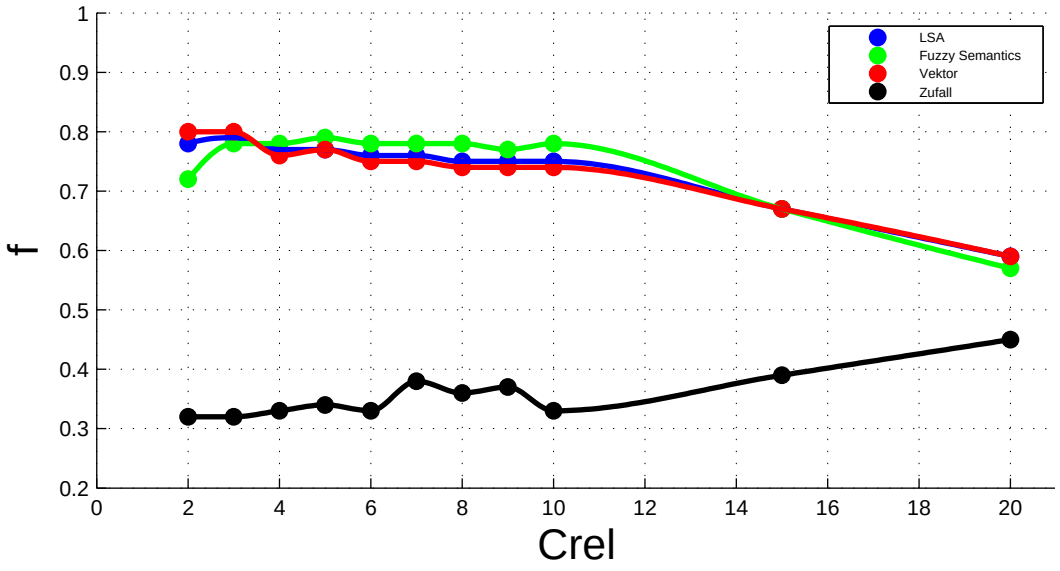

Abbildung 5.7: Grafische Darstellung der Evaluationsergebnisse, rein objektreferenzbasierter Ansatz, gemitteltes f

Wie erwartet ist die Anzahl an relevanten Symbolen für rein objektreferenzbasierte Berechnungen am geringsten. Die Anzahl der Symbole im gemischten Verfahren so wie im rein wortbasierten Verfahren unterscheiden sich kaum. Bei niedrigen Werten von C_{rel} stehen im gemischten Verfahren geringfügig mehr Symbole zum Tracking zur Verfügung, da oft vorkommende Objektgruppenbezeichnungen („chair“) disambiguiert werden (z.B. „`chair_06`“). Derselbe Effekt reduziert die Anzahl der Symbole, die bei hohen C_{rel}-Werten übrig bleiben.

Die Tabelle zeigt, dass im gemischten so wie im rein wortbasierten Verfahren eine Verdopplung von C_{rel} ungefähr einer Halbierung der Symbole entspricht. Somit fällt schon durch die Verdopplung von zwei auf vier Kontexte die Anzahl der relevanten Symbole in beiden Gruppen um ca. 300. Abhängig von dem Anwendungsgebiet der Themenerkennung kann dies nicht wünschenswert sein, je nachdem ob eine große Anzahl klassifizierter Symbole in Folgeprozessen benötigt wird. In diesen Fällen können niedrige Werte für C_{rel} erstrebenswert sein. Die Qualität der Klassifikation der selten vorkommenden Symbole ist dann allerdings – eben aufgrund ihres seltenen Vorkommens – fraglich.

History Die Tabelle 5.19 gibt einen Eindruck, wie oft bei den jeweiligen Verfahren und Werten für C_{rel} das Thema einer Äußerung nur durch Zugriff auf Information über das Thema der vorigen Äußerungen gefunden werden konnte, üblicherweise weil keine themenanzeigenden Symbole in der jeweiligen Äußerung vorhanden waren.

Die Werte in der Tabelle wurden über den drei Typen von semantischen Räumen gemittelt. Diese Vorgehensweise schien angebracht zu sein, da sich in keinem Fall die drei Werte um mehr als 0.6% voneinander unterschieden. Die Abweichungen ergaben sich aufgrund des Umstands, dass nur positive Themensummen als themenkennzeichnend bewertet wurden (vgl. Abschnitt 3.6). In Fällen, in denen keine positive Verbindungsstärkensumme der Symbole gefunden werden konnte, wurde automatisch

das letzte Thema angegeben. Die geringen Schwankungen deuten darauf hin, dass der größte Anteil an ausschließlich *history*-basierten Themenresolutionen jedoch die Fälle waren, in denen aufgrund mangelnder themenanzeigender Symbole das letzte Thema getrackt wurde.

C_{rel}	2	3	4	5	6	7	8	9	10	15	20
wortbasiert	69	66	64	63	60	58	56	54	53	47	39
gemischt	73	70	66	63	60	57	55	53	52	39	25
Objektreferenzen	53	51	48	46	44	42	40	38	37	25	14

Tabelle 5.19: Prozentualer Anteil von Themenresolutionen ohne *history*-Zugriff in Abhängigkeit von C_{rel}

Die Tabelle zeigt wie erwartet, dass die Anzahl der durch *history*-Zugriff klassifizierten Äußerungen mit steigendem C_{rel} stark abnimmt, wobei bei hohen C_{rel}-Werten und gerade im Fall des rein objektreferenzbasierten Trackings die *history*-basierten Klassifikationen die Anzahl der „echten“ Klassifikationen stark überschreitet. Dieser Umstand relativiert in diesen Fällen innerhalb gewisser Grenzen die Aussagekraft des Evaluationsverfahrens, wie auch schon durch die Berechnung der unteren Grenzen festgestellt wurde. Der Grund dafür liegt in der Bevorzugung von homogenenen Klassifikationen durch das Evaluationsverfahren: Der manuellen Annotation liegen Themen zugrunde, die sich fast immer über mehrere Äußerungen erstrecken. Ein Klassifikationsprozess, der die Tendenz hat, aufeinanderfolgende Äußerungen einem einzelnen Thema zuzuordnen, würde somit prinzipiell als besser bewertet werden, als ein solcher, der häufige Themenwechsel anzeigt. In Folge der Reduktion der Menge von Symbolen durch hohe Werte von C_{rel} werden immer weniger Äußerungen eigenständig, sondern über die *history* klassifiziert. Aus diesem Grund steigt die Homogenität der Klassifikation, was zu besseren Klassifikationsergebnissen führt, die aber nicht unbedingt einer „objektiv“ besseren Klassifikation entsprechen. Wie geschildert ist dieser Effekt durch die Berechnung einer *baseline* handhabbar. Man sollte nur im Hinterkopf behalten, dass gute Ergebnisse im Bereich von niedrigen C_{rel}-Werten eine größere Aussagekraft haben.

Betrachtet man die Ergebnisse im Bereich niedriger Werte für C_{rel}, so ist trotzdem ein verhältnismäßig hoher Anteil an *history*-basierten Klassifikationen zu bemerken. Dies erklärt sich aus der Existenz von Filler-Phrasen, die in jeder natürlichsprachlichen Interaktion häufig vorkommen.

Weiterhin interessant sind in diesem Zusammenhang die näherungsweise linearen Kurvenverläufe in Abhängigkeit von C_{rel}, die Symbolverteilungsprinzipien in dem Korpus nahelegen. Diese konnten aber in unabhängigen Untersuchungen mit Hilfe von Funktionsanpassungen nicht bestätigt werden[17].

[17]Die Funktionsanpassungen geschahen mit Hilfe des Altmann-Fitters (URL: `http://www.gabrielaltmann.de/interest.htm`). Herzlichen Dank an Alexander Mehler für die Durchführung der Analyse.

5.3.3 Diskussion

Nach Darstellung der Evaluationsergebnisse möchte ich einige Resultate zusammenfassen. Die beiden wohl wichtigsten Erkenntnisse sind, dass situiert-multimodale Themenerkennung im Kontext von Mensch-Roboter-Kommunikation trotz der großen Schwierigkeiten (im Vergleich zu Themenerkennung auf großen Datenkorpora) prinzipiell möglich ist, aber dass zur Bewältigung dieser Aufgabe multimodal-situierte Information – in diesem Fall in der Form der Auflösung von Objektreferenzen – benötigt wird. Aus diesem Umstand lässt sich schließen, dass überwacht lernende Themenerkennungsansätze aufgrund ihrer Schwierigkeit, Objektreferenzen – oder allgemein gesprochen: situative Gegebenheiten – in die Trainingsmenge mit aufzunehmen[18] wesentliche Nachteile im Vergleich zu unüberwacht lernenden Verfahren haben. Diese Erkenntnis wiederum stützt die Forderung nach dem Einsatz dynamischer Verfahren, die ja eine dieser Arbeit zugrundeliegende Basisprämisse ist.

Eine weitere Erkenntnis ist die, dass sich die Frage nach dem besten semantischen Raum nicht ohne weiteres beantworten lässt. Zwar ist der Rieger-Raum in bestimmten Fällen den anderen Räumen unterlegen, jedoch erzielen alle drei Räume unter bestimmten Bedingungen überraschenderweise sehr vergleichbare Ergebnisse. Grundsätzlich ist zumindest im gemischten Verfahren die LSA den anderen beiden Verfahren und im Speziellen dem einfachen Vektorraum überlegen, als dass sie auch bei geringeren Werten für C_{rel} gute Ergebnisse liefert, was zu einem größeren Informationsgehalt in den Themenclustern führt, da mehr Symbole Themen zugeordnet sind.

Den *online*-Experimenten (siehe Abschnitt 6.6) wurde jedoch der semantische Raum nach Rieger zugrundegelegt, da die Parametrisierung der LSA im Kontext sehr kleiner Räume/geringer Trainingsdatenmengen zunehmend problematisch wird. Weiterhin hatten Vorversuche gezeigt, dass im Vergleich zu dem einfachen Vektorraummodell durch den Fuzzy Semantics-Raum eine stärkere paradigmatische Assoziation thematisch verwandter Symbole stattfindet.

Eine wichtige Frage in Bezug auf eine mögliche *online*-Implementierung ist die Frage nach der Datengrundlage. Eindeutig ist, dass multimodale Information in Form von Objektreferenzen von großem Nutzen sein kann. Die Verwendung weiterer Symbole – so z.B. Ortskennzeichnungen oder Symbolen, die den emotionalen Zustand des Kommunikationspartners kennzeichnen – kann mit großer Wahrscheinlichkeit fruchtbare Resultate bringen, wurde aber im Rahmen dieser Arbeit nicht berücksichtigt. Ob Worte mit in die Verarbeitung aufgenommen werden sollen ist dagegen eine nichttrivial zu beantwortende Frage. Zwar können anscheinend mit dem Weglassen von Worten aus der Trainingsmenge nicht nur wesentlich bessere Resultate erzielt werden, sondern es konnte auch – wie ich in Abschnitt 5.4.3 darlegen werde – das dynamische Clustern gravierend verbessert werden. Allerdings gehen in diesem Kontext wichtige Informationen über die Themenzugehörigkeit von Worten zu Themen verloren, die im Einzelfall relevant für das Tracking oder für Themeninformation verarbeitende Module sein kann. Im folgenden Abschnitt 5.4.1 wird eine Strategie diskutiert, die möglicherweise die Vorteile beider Ansätze miteinander kombiniert.

[18] Natürlich kann auch durch den Benutzer eines Roboters ein überwacht lernendes Verfahren trainiert werden. Dies hätte jedoch den Nachteil, dass vom Benutzer ein nicht zumutbarer Aufwand betrieben werden müsste. Denkbar wären gemischte Ansätze aus explizitem und implizitem Lernen; ein solcher Ansatz wird in dem folgenden Kapitel geschildert.

Bevor ich mich dieser und anderen Fragestellungen zuwenden will, möchte ich jedoch noch ein Problem ansprechen, welches der vorliegenden Untersuchung als zugrundeliegend angesehen werden kann: Aufgrund des speziellen Kontextes (*robot home tour*), in dem das BITT-Korpus aufgenommen wurde, stellt sich die Frage, ob die gewonnenen Ergebnisse auf generische situierte Mensch-Roboter-Kommunikation anwendbar sind. Fraglich ist insbesondere die starke Themengebundenheit der Objekte.

Meiner Ansicht nach kann in den meisten Fällen diese Frage mit „ja“ beantwortet werden, da haushaltszentrierte Handlungen wie Abwaschen, Putzen etc. – die vermutlich die grundlegenden Themen in situierter Mensch-Roboter-Kommunikation darstellen werden – stets stark objektbezogen sind. Eine endgültige Antwort auf diese Frage lässt sich jedoch bei dem momentanen Wissensstand nicht geben, da unklar ist, welcher Art zukünftige Mensch-Roboter-Kommunikationen sein werden.

5.4 Ansatzpunkte

In diesem Abschnitt werde ich einige Punkte ansprechen, die über die mit den *offline*-Experimenten in Erfahrung zu bringenden Informationen hinausgehen. Im Wesentlichen handelt es sich dabei um Einzelansätze, die in späteren Kapiteln aufgegriffen und fortgeführt werden. Die Fälle, in denen dies nicht geschieht, werden zu einem großen Teil in dem vorliegenden Kapitel genauer untersucht.

Die geschilderten Experimente zeigen, dass automatische Themenerkennung auf Robotersystemen – sofern sie mit situiert-multimodaler Information unterstützt wird – zu verwertbaren Ergebnissen führen kann. Jedoch wurden aus verschiedenen Gründen in den Experimenten bestimmte Vereinfachungen gemacht und bestimmte Aspekte konnten nicht oder nur eingeschränkt untersucht wurden. Beispiele dafür sind:

1. Die Vorverarbeitungen erfolgten manuell und nicht auf einem Robotersystem.
2. Es wurde eine feste Themenanzahlsbegrenzung eingeführt (statisches Clustern).
3. Die Berechnungen erfolgten nicht inkrementell, wie bei einem *online*-System, sondern es wurde stets mit der Menge der restlichen Dialoge trainiert. Allgemeiner:
4. Die Experimente erfolgten insgesamt *offline*.
5. Bei der Segmentierung handelte es sich um eine „perfekte“, manuelle Segmentierung.

Punkt 1. kann als Kritikpunkt angeführt werden, da Fehler in der Vorverarbeitung, die auf einem Robotersystem entstehen würden, nicht oder nur eingeschränkt auftreten konnten. Auf diese Weise konnte die Robustheit des Systems gegenüber Fehlern z.B. in der Sprachverarbeitung nicht getestet werden. Es ist erwartbar, dass gerade Sprachverarbeitungsfehler bzw. Fehler in der Auflösung von Objektreferenzen dem Themenerkennungssystem ernsthafte Schwierigkeiten bereiten können, da es diese als Hinweise nicht nur für die Erkennung von Themen, sondern auch für das unüberwachte Training benötigt. Der Umstand, dass gerade Objektreferenzen starke Hinweise auf

das Thema liefern, ist allerdings von Vorteil, da diese in einem situiert arbeitenden, multimodalen Dialogsystem mit großer Wahrscheinlichkeit sehr robust verarbeitet werden können: So kann z.B. ein Fehler in der Spracherkennung, der zu einer fehlerhaften Objektzuordnung führen würde, sowohl rein sprachlich durch eine Rückfrage, aber auch zusätzlich durch visuelle Objekt- oder Gestenerkennung korrigiert werden.

Punkt 2. geschah wie beschrieben, um eine bessere Vergleichbarkeit der Ergebnisse zu gewährleisten. In Abschnitt 5.4.2 werde ich eingehend die Möglichkeiten dynamischer Segmentierung in Bezug auf das BITT-Korpus diskutieren. Die *online*-Experimente (Abschnitt 6.6) wurden mit echt dynamischer Segmentierung durchgeführt und zeigten, dass zumindest im Kontext kurzer Kommunikationssequenzen dynamisches Clustern möglich ist.

Eine inkrementelle Berechnung (Punkt 3.) konnte in der *offline*-Evaluation nicht angegangen werden, da inkrementelles Themenlernen zu sich stets verändernden Themen(clustern) geführt hätte, die eine Evaluation schwierig bis unmöglich gemacht hätten. Da bei dem gewählten Verfahren aber stets der Zustand des Systems nach einem gewissen Trainingsprozess simuliert werden konnte, halte ich diesen Punkt für unkritisch. Dass inkrementelle Themenberechnung prinzipiell möglich ist, zeigen wiederum die *online*-Experimente. Ähnlich verhält es sich mit Punkt 4., bei dem ich zusätzlich auf die Beschreibung des *online*-Systems verweisen möchte, in der Details der Implementierung dargelegt werden.

Kritisch und im Zusammenhang mit Punkt 1. zu sehen ist die mögliche Kritik (Punkt 5.), dass die durchgeführte Segmentierung quasi-optimal ist. Eine genauere Diskussion dieses Problems findet sich in diesem Kapitel in Abschnitt 5.4.3.

Eine letzte wichtige Fragestellung, die über die beschriebenen Vereinfachungen der *offline*-Experimente hinausgeht, ist natürlich die der Steigerung der Qualität der Ergebnisse in Kombination mit einer Optimierung der von dem Themenerkennungssystem gelieferten Information für nachfolgende Prozesse. Ich möchte mit der Diskussion dieses Punktes fortfahren. Anschließend wird auf die beschriebenen Probleme des dynamischen Clusterns und der dynamischen Segmentierung eingegangen.

5.4.1 Gemischte Modelle als Ausweg?

Wie beschrieben wäre es von großem Vorteil, die Evaluationsergebnisse des rein objektreferenzbasierten Verfahrens in Kombination mit den zusätzlichen Informationen des gemischten Verfahrens zu kombinieren.

Ich möchte an dieser Stelle einen Ansatz präsentieren, der dies mit großer Wahrscheinlichkeit zu leisten in der Lage ist: Trainiert man ein Themenerkennungsverfahren wie gewohnt ausschließlich mit Objektreferenzen, so lassen sich in dem anschließenden Trackingprozess die in den Tabellen 5.15, 5.16 und 5.17 angegebenen Ergebnisse erreichen. Die Vermutung war, dass die Beschränkung auf Objektreferenzen den Clusteralgorithmus entlastete, so dass sauberere Themencluster entstehen konnten. Weiterhin spielt wahrscheinlich auch das Fehlen von nicht-themenzugeordneten Worten, die nicht durch die Stopliste gelöscht wurden, eine Rolle, zumal vermutlich nicht-themenzugeordnete Objekte in den Monologen selten waren.

Um die fehlende Information über die Zugehörigkeit von Nicht-Objektreferenzen zu gewinnen, kann auf zwei verschiedene Weisen vorgegangen werden: Entweder clustert man einen Raum, der sowohl Objektreferenzen als auch Wortlemmata enthält, nur

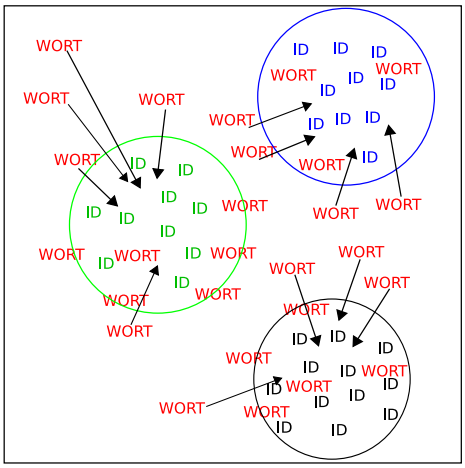

Abbildung 5.8: Gemischtes Verfahren Variante 1: Ein gemischter Raum (IDs/Worte) wird nur nach IDs geclustert, die Worte werden anschließend den Clustern zugeordnet

anhand der Objektreferenzen (vgl. Abbildung 5.8), oder man berechnet in einem ersten Schritt einen rein-objektreferenzbasierten Raum und in einem separaten Schritt einen gemischten. Dieser zweite Raum bzw. der vollständige Raum aus dem ersten Ansatz kann dann dazu verwendet werden, wie gehabt die Anbindung der Nicht-Objektreferenzen zu den jeweiligen Themenclustern zu berechnen, wobei an dieser Stelle allerdings nur die durchschnittlichen Distanzen zu den Objektreferenzsymbolen verwendet werden können (vgl. Abbildung 5.9).

In einem Experiment wurde die Leistungsfähigkeit dieses Ansatzes untersucht. Dazu wurden zwei verschiedene Räume berechnet – ein rein objektreferenzbasierter Raum und ein gemischter (Variante 2). Diese Doppelberechnung benötigt zwar mehr Rechenzeit als die eines einzelnen Raumes, aber es bestand bei Variante 1 die Unsicherheit, ob bei den semantischen Räumen nach Rieger oder bei der LSA in einem Raum auf der Basis von Worten und Objektreferenzen, bei dem ausschließlich die Objektreferenzen geclustert werden, dieselben Cluster entstehen würden wie in einem Raum, in dem nur Objektreferenzen enthalten sind. Dies erklärt sich durch die Beeinflussung der Position der Objektreferenzen durch die Wortvektoren in diesen beiden Räumen, selbst, wenn die Wortvektoren nicht geclustert werden. Durch Verwendung von Variante 1 konnte gewährleistet werden, dass ein Themenerkennungsvorgang tatsächlich dieselben Ergebnisse liefern würde, die in den Tabellen 5.15, 5.16 und 5.17 erzielt wurden.

Zur Berechnung der Themencluster wurde ein rein objektreferenzbasierter LSA-Raum mit drei Kontexten und manueller Segmentierung gebildet, da dieser in den geschilderten Versuchen die besten Ergebnisse erzielte. Anschließend wurden drei Räume gebildet, die zur Ermittlung der Verbindungsstärken der Wortsymbole herangezogen wurden. C_{rel} wurde in diesem Fall auf 2 gesetzt, um nur die seltensten Worte auszuschließen.

Als Räume wurden wiederum ein einfacher Vektorraum, ein semantischer Raum nach Rieger und ein LSA-Raum – alle entropie-gewichtet – verwendet. Zur Berech-

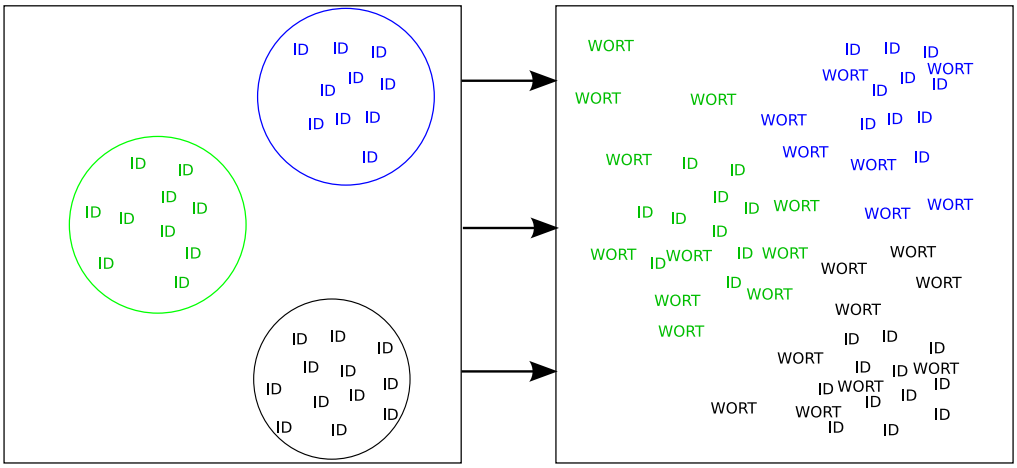

Abbildung 5.9: Gemischtes Verfahren Variante 2: Ein reiner Raum (IDs) wird geclustert, die Worte eines gemischten Raumes werden anschließend den geclusterten IDs zugeordnet

nung dieser Räume wurden jedoch nicht wie in den Versuchen mit gemischt aus Wortsymbolen und Objektreferenzen bestehender Datenbasis (Tabellen 5.12, 5.13 und 5.14) die objektreferenzierenden Worte durch die Objekt-ID überschrieben, sondern es wurden sowohl die Objekt-IDs als auch die Lemmata der referenzierenden Nominalphrasen mit trainiert. Die Löschung der Worte geschah – wie in Abschnitt 5.3.2 beschrieben – um den Clustervorgang zu erleichtern. Da bei der verwendeten Variante 2 der Clustervorgang der Worte entfällt und auf diese Weise wieder die Information der Zugehörigkeit von Objekttypen zu Themen ermittelt werden kann, wurde diese Vorgehensweise gewählt.

Um die anschließende Berechnung der Anbindungsstärken der Wortsymbole an die jeweiligen Themencluster zu evaluieren, wurde wie folgt vorgegangen: Für jedes Symbol wurde das Thema mit der stärksten Anbindung bestimmt. Als Referenz wurde für jedes Wort der Liste manuell die Gruppe an möglichen Themen ermittelt, zu denen das Wort thematisch passte. Die diese Arbeit ausführende Person hatte fundierte Kenntnisse von dem BITT-Korpus und war somit in der Lage zu entscheiden, wie die Zuordnung im Kontext der Situationen der BITT-Monologe sinnvoll zu treffen war.

In einem nächsten Schritt wurde für jede der automatischen Zuordnungen geprüft, ob das ermittelte, am stärksten verbundene Themencluster in der manuell erstellten Liste von relevanten Themen zu finden war. Im Fall einer leeren Liste wurde das Wort aus der Evaluation herausgenommen; oft handelte es sich dabei um Worte, die sinnvollerweise auf der Stopliste hätten verortet sein sollen, dies aber nicht waren.

In Tabelle 5.20 entsprechen diese Worte der Spalte „kein Thema“. Wie aus der Tabelle ersichtlich ist, liegt der ungefähre Anteil von korrekten Zuordnungen unabhängig vom verwendeten semantischen Raum bei ca. 80%.

Aufgrund des verhältnismäßig gutmütigen Evaluationsprozesses, der nur verlangt, dass sich das automatische Ergebnis innerhalb einer Liste von manuell als relevant gekennzeichneten Themen befindet, überrascht dieses eher mäßige Ergebnis. Trotzdem zeigt Tabelle 5.20, dass sich in den berechneten Anbindungsstärken von Worten an

Raum	#korrekt	#falsch	#kein Thema	% korrekt
Einf. Vektorraum	272	70	178	80
Rieger	273	69	178	80
LSA	264	78	178	77

Tabelle 5.20: Evalutionsergebnisse - nachträgliche Berechnung der Anbindungsstärken von Wortsymbolen

objektreferenzbasierte Themencluster Information verbirgt, aus der – unter Vorbehalt – Nutzen gezogen werden kann, ohne die Qualitätseinbußen hinnehmen zu müssen, zu denen das gemischt wortsymbol/objektreferenzbasierte Verfahren gegenüber dem rein objektreferenzbasierten führt. Außerdem wird auf diese Weise ermöglicht ohne größeren Aufwand Fragen der Art „Welche Arten von Objekten gehören zur Spüle?“ von dem Robotersystem beantworten zu lassen, was bei dem bisherigen Ansatz (vgl. Abschnitt 5.3.2) nur mittelbar möglich war.

In einem weiteren Schritt könnten auf diese Weise auch weitere Modalitäten sinnvoll berücksichtigt werden, ohne die Trackingqualität zu verschlechtern. Ein gutes Beispiel wäre die Einbindung von Emotionalität, für die im Kontext von BIRON schon Grundlagen geschaffen wurden (Hegel u. a., 2006). Zweifellos zeigt eine bestimmte Emotion nur in sehr geringem Maße bzw. unspezifisch ein Thema an, aber die (umgekehrte) Information, dass Gesellschaftsspiele mit Freude, der Arbeitsbereich aber mit Ernsthaftigkeit (oder sogar Aggression durch die Störung durch den Roboter) verbunden ist, kann von großem Nutzen sein. Ob und wie sinnvoll sich solche Prozesse realisieren lassen ist aber im Rahmen dieser Arbeit nicht zu beantworten, sondern kann erst im Rahmen zukünftiger Forschungen im Bereich der *interaction robotics*, die über fortgeschrittenere Roboteragenten verfügen, untersucht werden.

5.4.2 Dynamisches Clustern

In den im Kapitel 5.3.2 geschilderten Untersuchungen wurde der agglomerative Clusteralgorithmus künstlich auf ein Limit von maximal 10 zu erkennenden Themen beschränkt. Da das Fernziel dieser Arbeit ein echt dynamisches Clustern vorkommender Symbole in Themen umfasst, stellt sich die Frage nach der Erreichbarkeit dieses Zieles.

Wie in Abschnitt 3.4.6 dargestellt wurde, unterteilt das verwendete Distanzmaß der Korrelation einen semantischen (Vektor-)Raum in korrelierte und un- bzw. antikorrelierte Vektoren. Es wäre naheliegend, bei null oder in der Nähe von null eine Grenze zu ziehen, die nicht von dem agglomerativen Clusteralgorithmus überschritten werden darf.

Eine andere Vorgehensweise wäre eine Analyse der Distanzen zwischen Vektormengen, die der Clusteralgorithmus zusammenfügt[19]. Kommt es in dem Verlauf dieser Distanzen zu erkennbaren Sprüngen, ist dies ein Hinweis auf eine aus den Daten ermittelbare Grenze, an der sich ein Themenerkennungsprozess automatisch beenden könnte.

Um beide Hypothesen zu untersuchen, wurden für die manuelle Segmentierung

[19]vgl. dazu Abschnitt 3.5

Kurven der Distanzen von Vektormengen erstellt, die der Clusteralgorithmus im jeweiligen Schritt zusammenfügt. Sie sind in der Tabelle 5.21, so wie gesondert in den Grafiken 5.10, 5.11 und 5.12 aufgeführt. Die Kurven bzw. Tabellen sind nur für die letzten 50 bzw. 40 Clusterschritte[20] angegeben, um so eine bessere Übersichtlichkeit zu gewährleisten – in den nicht dargestellen Bereichen finden sich keine Sprünge, sondern ausschließlich kontinuierliche Verläufe.
Als Basisraum wurde wie zu der Berechnung des gemischten Ansatzes auf Seite 102 ein LSA-Raum mit einem C_{rel} von 3 verwendet. Nicht zuletzt aufgrund der Verwendung einer LSA sind die Kurven leicht nach oben verschoben und enthalten – abgesehen von Grafik 5.12 – keine Werte im negativen Bereich. Die einzige Kurve, in der sowohl negative Werte als auch ein per Auge eindeutig erkennbarer Knick in der Kurve vorliegen, ist Grafik 5.12. Der Sprung ist bei exakt acht Clustern angesetzt, einer Zahl, die der durchschnittlichen Themenanzahl in den manuellen Annotationen entspricht.

Was bedeutet dieses Ergebnis für die Aussicht auf eine echt dynamische Themengruppierung? Echt dynamische Themenclusterung lässt sich augenscheinlich nur bei optimaler Segmentierung und einer optimalen Symbolwahl realisieren. Eine Grenze bei einem Korrelationswert von 0 zu ziehen, schlägt ansonsten – zumindest bei der LSA – fehl. Ein Ausweg bestünde darin, einen willkürlich gewählten Wert knapp über 0 zu wählen; eine solche Vorgehensweise ist aber potentiell unbefriedigend, da sie fast immer suboptimale Ergebnisse liefert.

An dieser Stelle muss aber auch angemerkt werden, dass in einfachen Szenarien bei guter bis sehr guter Segmentierung dynamisches Clustern durchaus realistisch ist. Gezeigt wird dies nicht zuletzt in den *online*-Experimenten in Kapitel 6.

5.4.3 Probleme dynamischer Segmentierung

Wie geschildert ist die *offline*-Evaluation mit Hilfe einer Segmentierung durchgeführt worden, die aus der manuellen Themenannotation des BITT-Korpus abgeleitet wurde. Aus diesem Grund kann die Segmentierung als optimal angesehen werden. Als Resultat fand eine Evaluation des Themenerkennungssystems unter perfekten Bedingungen[21] bezüglich aller sensorischen Prozesse statt – die Spracherkennung „funktionierte“ optimal, die Auflösung von Objektreferenzen ebenso und es kam zu keinen Fehlern durch einen – in einem *online*-System zweifellos ebenfalls durch multimodale Daten gestützten – Segmentierungsprozess.

Diese Abstraktion war im Kontext der Experimente wünschenswert, da so die Kernalgorithmen der Themenerkennung gezielt getestet werden konnten. Auf der anderen Seite wären natürlich Einblicke in die Leistungsfähigkeit des Systems unter realen Bedingungen nützlich.

Die Problematik fehlerhafter Sprach- und Objekterkennung wurde weiter oben (Abschnitt 5.4) kurz angesprochen, sie konnte aber im Rahmen dieser Arbeit nicht mehr untersucht werden.

Dynamische Segmentierung findet in den *online*-Experimenten statt, allerdings wurde diesen Experimenten ein verhältnismäßig einfaches Szenario zugrundegelegt,

[20] Es handelt sich um die letzten Schritte, da der Clusteralgorithmus agglomerativ, also „*top-down*“ vorgeht. Bei Schritt eins wären alle Vektoren zu einem Cluster zusammengefasst.

[21] Natürlich treten unweigerlich auch Fehler in den manuellen Annotationen auf, so dass das Wort „perfekt“ als relativ zu betrachten ist.

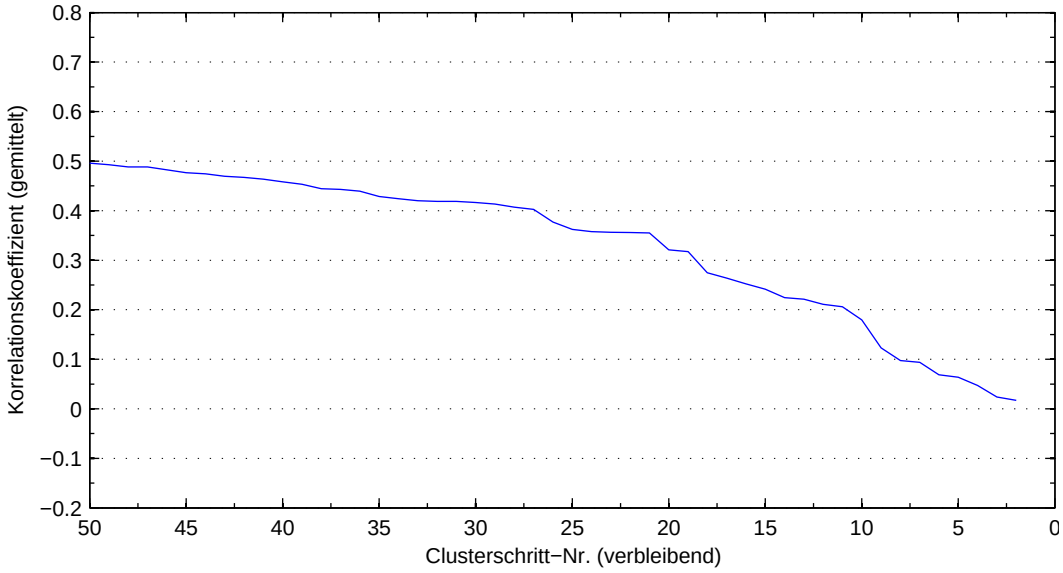

Abbildung 5.10: Clustergrenzen bei Wortlemmata (LSA, C_{rel}=3)

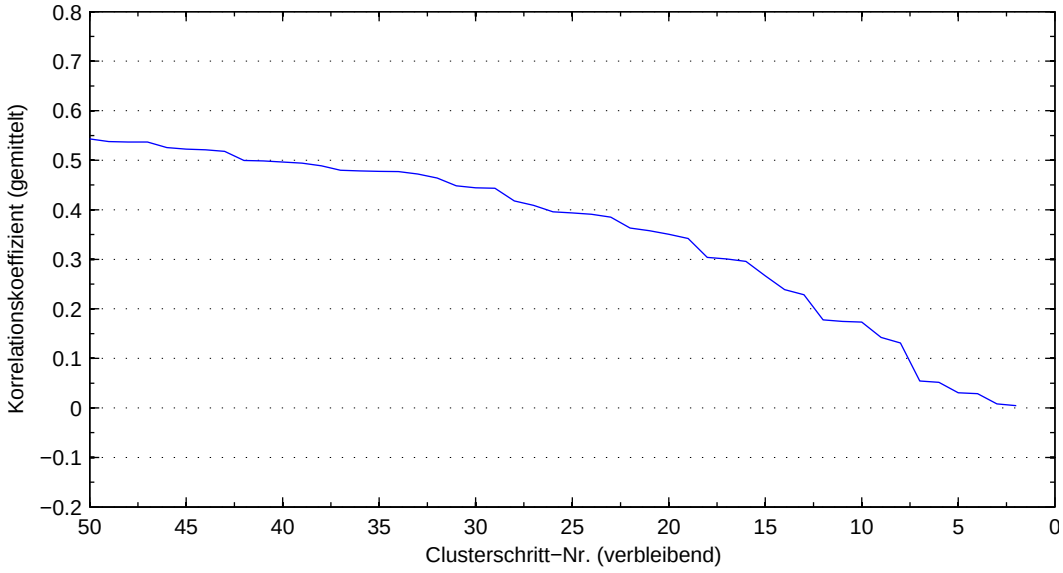

Abbildung 5.11: Clustergrenzen bei Wortlemmata und Objektreferenzen (LSA, C_{rel}=3)

Schritt	Clusterlimit (Korrelation)		
	wortbasiert	**gemischt**	**Objektref.**
40	0.4582	0.4963	0.7564
39	0.4534	0.4941	0.7368
38	0.4441	0.4887	0.7290
37	0.4429	0.4797	0.7284
36	0.4393	0.4783	0.7243
35	0.4285	0.4775	0.7160
34	0.4242	0.4771	0.7144
33	0.4202	0.472	0.6826
32	0.4188	0.464	0.6704
31	0.4187	0.4485	0.6630
30	0.4163	0.4442	0.6581
29	0.4135	0.4433	0.6514
28	0.4070	0.4180	0.6420
27	0.4024	0.4088	0.6274
26	0.3771	0.3960	0.6239
25	0.3621	0.3938	0.6139
24	0.3579	0.3909	0.6097
23	0.3561	0.3850	0.6059
22	0.3559	0.3630	0.5984
21	0.3552	0.3579	0.5790
20	0.3208	0.3506	0.5534
19	0.3173	0.3422	0.5509
18	0.2745	0.3038	0.5306
17	0.2637	0.3008	0.5024
16	0.2521	0.2957	0.4717
15	0.2414	0.2667	0.4556
14	0.2246	0.2385	0.4514
13	0.2214	0.2284	0.4008
12	0.2108	0.1776	0.3959
11	0.2061	0.1745	0.3792
10	0.1794	0.1733	0.2929
9	0.1228	0.1421	0.2874
8	0.0972	0.1310	0.1793
7	0.0936	0.0541	-0.0126
6	0.0688	0.0515	-0.0192
5	0.0635	0.0306	-0.0215
4	0.0473	0.0286	-0.0286
3	0.0236	0.0079	-0.0290
2	0.0171	0.0044	-0.0385

Tabelle 5.21: Clusterlimits (Korrelation) beim hierarchisch-agglomerativen Clustern auf dem BITT-Korpus

bei dem nur wenige Segmentierungsfehler auftraten. Eine Untersuchung der Auswirkungen fehlerhafter Segmentierung anhand des BITT-Korpus oder anhand von weitergehenden Experimenten scheint somit naheliegend zu sein.

Bevor jedoch die Frage nach der Qualität der Themenerkennung unter „realistischen“ Umständen beantwortet werden kann, muss zuerst die Frage gestellt werden, wie gute Ergebnisse eine solche Segmentierung überhaupt zu liefern in der Lage ist. Die endgültige Antwort auf diese Frage kann im Rahmen dieser Arbeit nicht gegeben werden, da die Qualität der Segmentierung vermutlich stark von den Fähigkeiten des künstlichen Kommunikators so wie der Art der Kommunikation abhängt. Allerdings lassen Forschungen im Bereich der Textsegmentierung, Sprechersegmentierung und Dialogsegmentierung den Schluss zu, dass qualitativ hochwertige Segmentierungen möglich sind. Eine genauere Diskussion dieses Punktes findet sich in Abschnitt 6.3.

Um einen ungefähren Eindruck der Leistungsfähigkeit des Themenerkennungssystems unter suboptimaler Segmentierung zu gewinnen, wurden in einem frühen Stadium der Arbeit analog zu den beschriebenen Versuchen Vorexperimente auf dem BITT-Korpus durchgeführt. Problematischerweise ist eine optimale Strategie zur Segmentierung auf dem Korpus aufgrund seiner monologischen Struktur wahrscheinlich keine optimale Strategie in realer Kommunikation. Letztendlich lässt sich jedoch hoffen, dass sich aufgrund der stärkeren Eingebundenheit des künstlichen Kommunikators in realer Kommunikation prinzipiell bessere Segmentierungsergebnisse erzielen lassen als auf dem Korpus. Aus diesem Grund wurde im Rahmen des Experiments versucht, eine gute – wenn auch nicht optimale – automatische Segmentierungsstrategie für den Korpus zu finden und die Auswirkungen auf die Evaluationsergebnisse zu beobachten.

Nützliche Hinweise dazu fanden sich in den in (Shriberg u. a., 2000) dargestellten Ergebnissen: formantenverlaufsbasierte Segmentierungsalgorithmen leisten gute Arbeit im Kontext von ununterbrochener, nicht-situierter Sprache, wie z.B. bei Radionachrichten. Zur Segmentierung von dialogischer Sprache – oder monologisch-situierter wie im BITT-Korpus – leisten pausenbasierte Algorithmen wesentlich bessere Arbeit. Die Segmentierung, die in den Vorexperimenten verwendet wurde, basiert auf der Kombination von zwei pausenbasierten Segmentierungsprozessen mit einer Wortwiederholungsanalyse. Als Themengrenzen erkannt wurden alle Pausen, die länger als 90% der in dem Monolog vorkommenden Pausen waren, oder die länger als 1.4 Sekunden waren. Um die Fehler durch „Denkpausen“ – also nicht-themenwechselnde Pausen – zu minimieren, wurden Pausen ignoriert, die zwischen Äußerungen lagen, die über ein oder mehrere identische Worte verfügten. Worte auf der Stopliste wurden dabei selbstverständlich nicht gewertet.

Aufgrund der geringen Aussagekraft und des Umfangs der Ergebnisse wird an dieser Stelle auf eine detaillierte tabellarische Darstellung verzichtet. Als Vorversuche konnten die Experimente jedoch die Ergebnisse der Experimente mit manueller Segmentierung vorwegnehmen: Auch in ihnen war der positive Effekt der Einbindung von Objektreferenzen klar erkennbar. Trotzdem waren die erreichten absoluten Ergebnisse eher schlecht, was die Abhängigkeit der Themenerkennung von der Segmentierung deutlich macht. Im Rahmen dieser Arbeit wurde jedoch auf eine weitere Optimierung der Segmentierung auf den *offline*-Daten verzichtet, da sie sich wie beschrieben grundsätzlich von einer *online*-Segmentierung unterscheidet; stattdessen wurde das System *online* implementiert und in diesem Rahmen eine Segmentierung ausprobiert.

Diese Vorgänge und resultierende Experimente werden im folgenden Kapitel beschrieben.

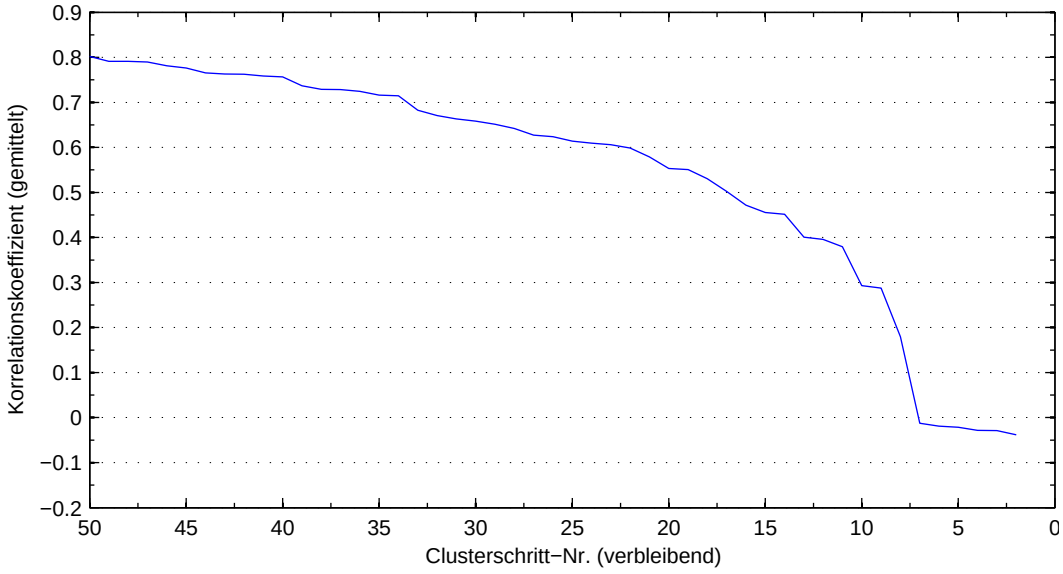

Abbildung 5.12: Clustergrenzen bei Objektreferenzen (LSA, C_{rel}=3)

6 Online-Implementierung und Evaluation

Nachdem die Funktionsfähigkeit des Themenerkennungsalgorithmus erfolgreich *offline* getestet worden war, bestand der nächste Schritt in der Implementierung eines Moduls zur Themenerkennung in die Hard- und Softwarearchitektur des Roboters BIRON.

Dazu war neben der Aufgabe der eigentlichen Integration – also der Anbindung an roboterinterne Ressourcen etc. – insbesondere die Frage nach einer passenden Segmentierung zu lösen: Klassische Textsegmentierungsansätze, wie sie z.B. bei (Purver u. a., 2006) oder (Hearst, 1997) zu finden sind, basieren auf einem Training des Segmentierungsalgorithmus anhand eines Korpus. So basiert das in (Hearst, 1997) dargestellte *TextTiling*-System z.B. auf der Berechnung von Textkohäsion in Trainingsdaten, die dann anhand der gewonnenen Information segmentiert werden. Obwohl es sich hierbei um einen *bootstrapping*-Ansatz handelt – das Verfahren wird unüberwacht trainiert, wobei keine Unterscheidung zwischen Trainings- und Testdaten vorliegt – ist eine gewisse Basismenge an Daten von Nöten.

Es ist davon auszugehen, dass im Rahmen weiterer Entwicklungen von Diskurssegmentierungsalgorithmen sehr gute Algorithmen und für eine Erweiterung auf allgemeine Anwendungsfälle geeignete Testdaten gewonnen werden können. In dieser Arbeit wurde jedoch ein anderer Ansatz verfolgt, der (fast) keine Trainingsdaten benötigt und somit – wie das Themenerkennungssystem im Idealfall auch – ab der ersten Roboter-Benutzer-Kommunikation funktionsfähig ist.

Ein weiterer Aspekt der Implementierung bestand in dem Entwurf einer modulinternen Struktur, welche insbesondere die Aufgabe des Trackings schnell durchzuführen ermöglichte. Da aber die Struktur des Themenerkennungsprozesses von Anfang an darauf ausgelegt war, trotz eines potentiell zeitintensiven Trainings ein schnelles Tracking zu gewährleisten, bereitete dieser Punkt bei der Implementierung nur geringe Schwierigkeiten.

Bevor jedoch die Details der Implementierung dargelegt werden, möchte ich auf die Hardware- und Softwarearchitektur des Roboters BIRON, so wie seine bisherigen Fähigkeiten eingehen.

6.1 BIRON

Bei dem Roboter BIRON (Bielefeld Robot Companion) handelt es sich um einen modifizierten Pioneer2 PeopleBot von ActivMedia (vgl. Abbildung 6.1). Diese Roboterbasis wurde speziell für Anwendungen entwickelt, die einen mobilen Roboter benötigen, der mit Menschen zu kommunizieren in der Lage ist. Ein Beispiel dafür sind automatische Museumsführer. BIRON kann kurze Zeit völlig autark über die eingebauten Akkumulatoren versorgt werden, die Laufzeit beträgt zwischen 30 Minuten

(bei Bewegung) und zwei Stunden.

6.1.1 Fähigkeiten des Systems

Um die Funktionsweise der einzelnen Module und Hardwarekomponenten einfacher ersichtlich zu machen, möchte ich an dieser Stelle eine kurze Übersicht über die Fähigkeiten des Systems geben. Dabei ist zu beachten, dass sich das System in einem steten Prozess der Fortentwicklung befindet und daher diese skizzenhafte Beleuchtung notwendigerweise unvollständig ist. Ich möchte darum nur kurz auf die Kernfähigkeiten eingehen.

Auf BIRON wurden große Teile des *home tour*-Szenarios verwirklicht. Der Roboter ist in der Lage, zwischen mehreren möglichen Kommunikationspartnern zu differenzieren und in dem Moment, in dem er angesprochen wird, sich auf einen Kommunikationspartner zu fokussieren. BIRON kann begrüßt werden, versteht es, wenn er verabschiedet wird und kann auf Anweisung Personen folgen, so wie ihm bekannte Orte selbständig wieder aufsuchen. Weiterhin kann er einfache Zeigegesten erkennen und mit Hilfe solcher so wie verbaler (z.B. sprachlich spezifizierter Farb-)Information Objekte erkennen und Ansichten derselben abspeichern. Wichtig ist natürlich auch BIRONs Fähigkeit u.a. mittels „gesprochener“ Sprache auf Benutzeraktionen angemessen zu reagieren.

6.1.2 Hardware

BIRON verfügt über verschiedene Sensorentypen, die die Kommunikation mit Menschen ermöglichen sollen:

Distanzsensoren Standardmäßig ist die Roboterbasis mit Ultraschallsensoren ausgestattet, die Kollisionen mit Objekten verhindern sollen. Diese wurden jedoch in BIRON durch einen frontal montierten SICK LMS200 Laserscanner ersetzt. Dieser verfügt über einen (rein horizontalen) Scanwinkel von 183°, so wie eine maximale Scanreichweite von 100m. Der Scanner kann 4.7 Scans pro Sekunde mit einer Auflösung von 0.5° durchführen, was zu 361 Einzelmessungen pro Scan führt.

Aufgrund der Montagehöhe von 30cm eignet er sich, Beinpaare – und somit aktuelle oder potentielle Kommunikationspartner – zu lokalisieren.

Audio BIRON verfügt über zwei AKG C 400 BL Grenzflächenmikrophone (106cm Höhe. 28.1cm Distanz). Diese sind speziell dazu geeignet, schräg von vorne/oben kommende Schallquellen gegenüber Hintergrundgeräuschen aufzunehmen. Die Stereomikrophone dienen sowohl zur Aufzeichnung von gesprochener Sprache von Interaktionspartnern so wie der Sprecherlokalisation (Hohenner u. a., 2003). Durch generische Anschlüsse ist es weiterhin möglich, BIRON z.B. mittels eines Funkmikrophons anzusprechen.

Video BIRON verfügt über zwei Kameras, die unterschiedlichen Zwecken dienen. Die auf 142cm Höhe (und somit auf der Spitze von BIRON) montierte Pan-Tilt-Kamera (Sony EVI-D31) dient insbesondere der Gesichts- und Objekterkennung. Sie

verfügt über einen horizontalen maximalen Öffnungswinkel von 48.8° und einen vertikalen maximalen Öffnungswinkel von 37.6°. Durch die Bewegung der Kamera auf der vertikalen Achse kann die Kamera um 100°, auf der horizontalen Achse um bis zu 25° bewegt werden.

Aufgrund des Umstands, dass BIRON im Betrieb versucht, das Gesicht des Kommunikationspartners mit dieser Kamera zu verfolgen, kann mit Hilfe dieses Sensors z.B. erkannt werden, ob eine sprechende Person BIRON ansieht (was zu einer Selektion als Gesprächspartner führt) oder nicht. Interessanterweise handelt es sich hierbei um eine zwei-Wege-Kommunikation, da erkennbar ist, ob eine und wenn, welche Person von BIRON „betrachtet“ wird. Eine weitere wichtige Funktion dieser Kamera ist das Scannen von Objekten, auf die der Kommunikationspartner gezeigt hat.

Die auf 95cm Höhe montierte Apple iSight-Kamera verfügt über einen maximalen Öffnungswinkel von 54.3°, die Kamera ist allerdings starr nach vorne ausgerichtet. Sie dient insbesondere der Erkennung von Zeigegesten des Kommunikationspartners. Diese sind meistens sprachbegleitend, so dass in diesen Situationen die Pan-Tilt-Kamera auf das Gesicht des Gegenübers ausgerichtet ist und keine Gesten identifizieren kann.

Eine weitere Anwendung für diese Kamera ist eine Personenerkennung, die ein Körpermodell des Kommunikationspartners beinhaltet (Schmidt u. a., 2006). Diese befand sich im Moment der Erstellung dieser Arbeit allerdings noch in der Entwicklung.

Zusätzlich verfügt BIRON über folgende **Anzeigeelemente**:

Touch Screen BIRON verfügt über einen Bildschirm, der im Wesentlichen verwendet werden kann, um den Systemzustand anzuzeigen. In der Kommunikation mit naiven Benutzern wird auf dem Bildschirm das „*robby face*“ angezeigt, welches durch die Darstellung eines *comic*-haften Gesichtes basale, simulierte Emotionen kommunizieren kann.

Der Bildschirm ist berührungssensitiv und kann somit auch als Eingabegerät verwendet werden. Dieses Feature wird allerdings bisher in der Mensch-Roboter-Kommunikation nicht verwendet, könnte aber z.B. die Objekterkennung durch manuelle Objektselektion unterstützen.

Stereolautsprecher Weiterhin kann BIRON mit Hilfe eines Sprachsynthesesystems über die integrierten Stereolautsprecher Sprache ausgeben. Im Rahmen von multimodaler Mensch-Roboter-Kommunikation wird von dieser Möglichkeit natürlich stark Gebrauch gemacht.

Aktuatoren BIRON besitzt zur Zeit keine Aktuatoren außer denen zur Fortbewegung, da bisherige Experimente noch keine von BIRON ausgeführten Objektmanipulationen vorsahen. Der Roboter ist in der Lage mit Hilfe von Rädern und Elektromotoren in der Roboter-Basis sich zu drehen und Personen zu folgen, oder aber selbständig bekannte Orte aufzusuchen.

6.1.3 Rechnerleistung

In die Basis wurden zwei PCs (PentiumII, 500mHz und 850mHz) integriert, die aber aufgrund der schnellen Entwicklung der Computersysteme mittlerweile etwas veraltet

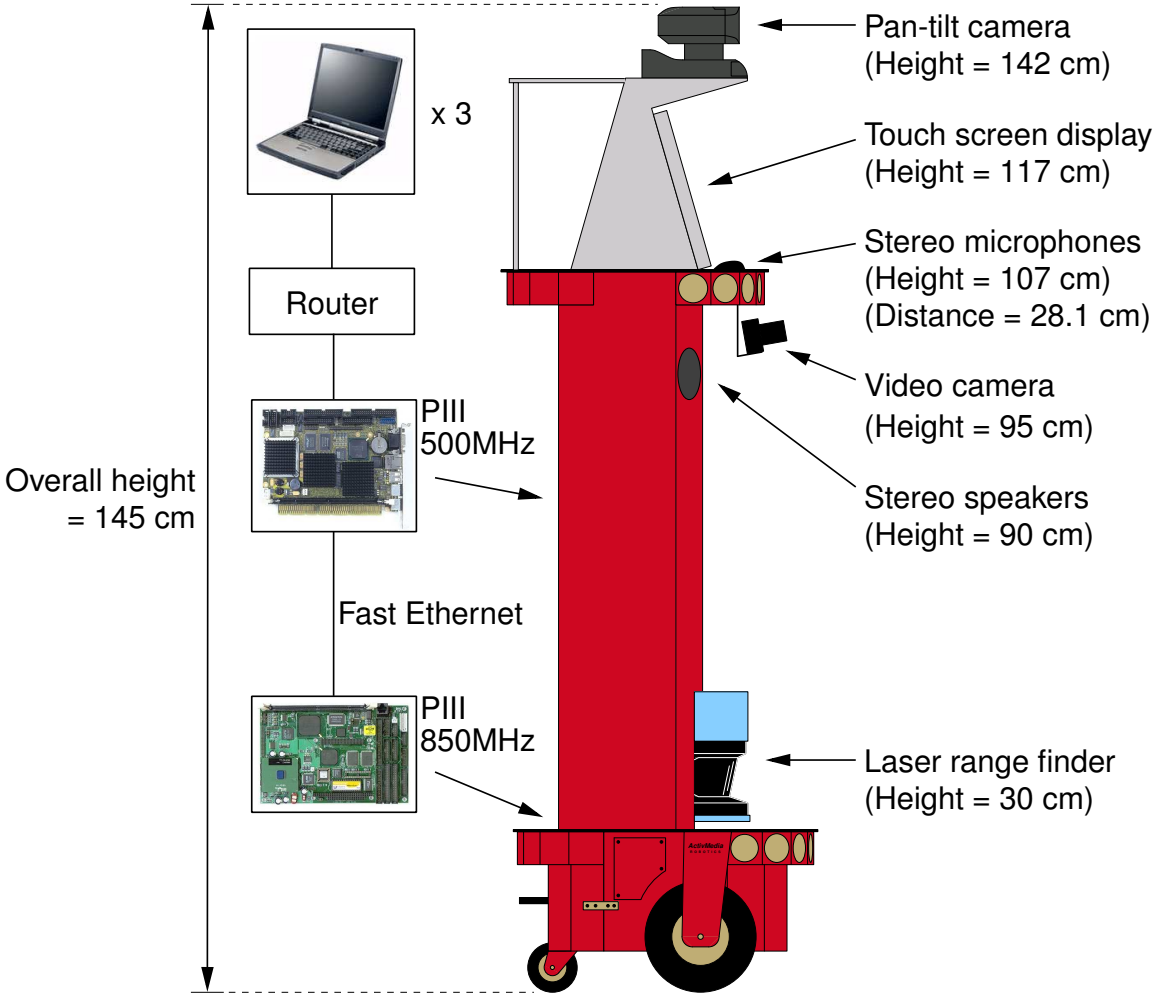

Abbildung 6.1: Schematische Darstellung von BIRON

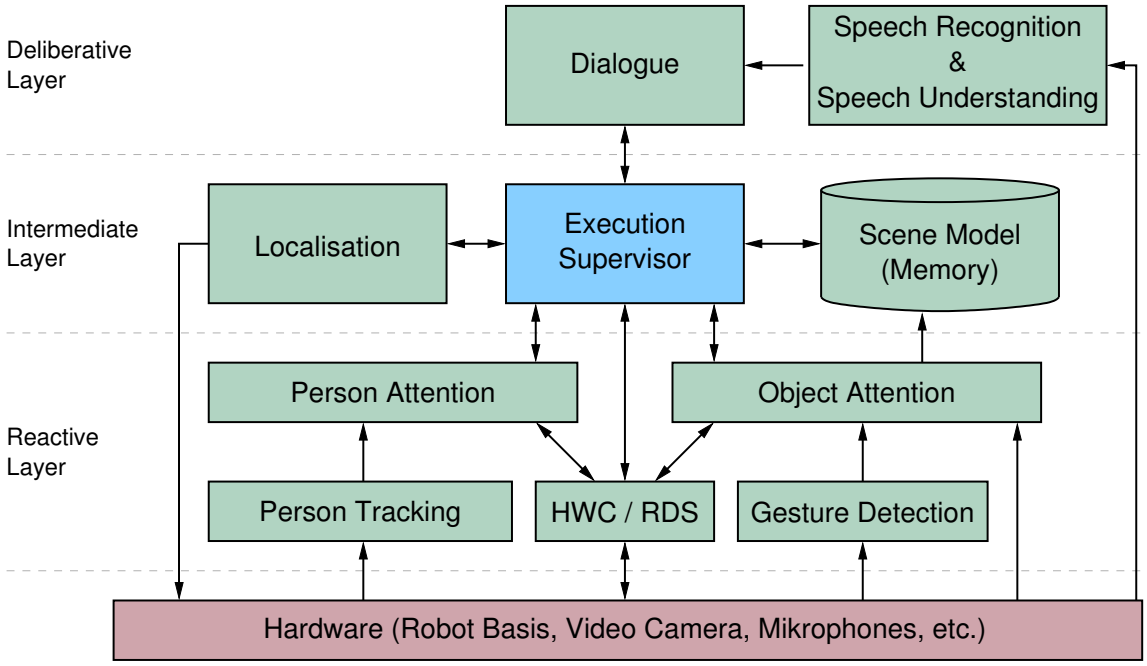

Abbildung 6.2: Drei-Schichten-Modell der Software

sind. Aus diesem Grund können an den Seiten von BIRON zwei Notebooks befestigt, so wie ein weiteres über ein WLAN angeschlossen werden. Das interne Netzwerk ist ein Fast-Ethernet-Netzwerk, welches die Rechner miteinander verbindet.

6.1.4 Softwarearchitektur

Um seiner Aufgabe – der situierten Kommunikation mit Menschen über eine anthropozentrische Umgebung – nachgehen zu können, wurde für BIRON eine komplexe Softwarearchitektur entwickelt.

Struktur Die für BIRON entwickelte Software beruht auf einer dreischichtigen Architektur (Haasch u. a., 2004) (Fritsch u. a., 2005), wie in Abbildung 6.2[1] vereinfacht[2] dargestellt. Dabei wird zwischen Komponenten auf der **reaktiven Ebene** – also solchen, die direkt auf sensorisches Input reagieren und es weiterverarbeiten, zwischen Komponenten auf der **Zwischenebene** – also insbesondere Komponenten, die halbintelligent arbeiten bzw. die die Zustände des Systems verwalten – und Komponenten auf der **deliberativen Ebene** – also Komponenten, die höher intelligente Steuerungsfunktionen übernehmen – unterschieden.

Im Folgenden möchte ich auf einzelne Komponenten eingehen, die für das Robotersystem von großer Relevanz sind. Ein Fokus dieser Darstellung wird aber auch auf der Verwendbarkeit der einzelnen Komponenten für die beiden Basisaufgaben der Themenerkennung – Tracking und Segmentierung – liegen.

[1] In der Abbildung wurden die Elemente in Englisch bezeichnet, da es sich in diesen Fällen um übliche Bezeichnungen handelt. Die Grafik wurde (Fritsch u. a., 2005) entnommen.

[2] Die Abbildung zeigt verschiedene Module in Gruppen (Kästchen) zusammengefasst. Eine genauere Aufschlüsselung wäre im Rahmen dieser Arbeit nicht zweckmäßig gewesen, daher wurde diese Darstellungsweise gewählt.

XCF Als Grundlage der Architektur dient das im Rahmen der Arbeitsgruppe Angewandte Informatik entwickelte **XML Communication Framework** (XCF) ((Wrede u. a., 2004) bzw. `http://xcf.sourceforge.net/` (homepage)). Es ermöglicht den Austausch und die Validierung von beliebigen XML-Dokumenten über Computernetzwerke, wobei die Dokumente auch Binärdaten (sog. BLOBs[3]) enthalten dürfen. Die Möglichkeit der Validierung erlaubt standardisierte Informationsübermittlung und ist die Grundlage für eine effiziente Modularisierung des Systems, die *rapid prototyping* ermöglicht. Da XCF auf jedem am Netzwerk beteiligten Computersystem Zugriff auf alle angebundenen Ressourcen zur Verfügung stellt, können Komponenten verteilt auf einer wechselnden Hardwarebasis ausgeführt werden, ohne an der Systemarchitektur starke Modifikationen vornehmen zu müssen.

Mit XCF ist es sowohl möglich, eine Broadcast/Listener, als auch eine verteilte Funktionsaufruf-Architektur oder eine gemischte Architektur aufzubauen. XCF unterstützt die Anbindung sowohl an Java, Matlab als auch C/C++.

Trotz der Entwicklung dieses Werkzeugs findet zwischen bestimmten Modulen immer noch Kommunikation in reinen Binärformaten statt, dies aber fast nur zwischen Applikationen, die eng miteinander verknüpft sind und auf demselben physikalischen Rechnersystem ausgeführt werden.

Im Folgenden möchte ich die einzelnen Softwarekomponenten von BIRON im Detail darstellen.

Execution Supervisor Das zentrale Steuerungsmodul ist der sog. *Execution Supervisor* (ESV) (Kleinehagenbrock, 2005) (Fritsch u. a., 2005). Dieses Modul besteht aus einem per XML-Datei konfigurierbaren Erweiterten Finiten Automaten (AFSM[4]). Einzelne Module können Nachrichten an einen Verarbeitungsstack schicken und so den Systemzustand ändern, um global für das System erkennbare Modifikationen vorzunehmen. Eine Änderung des Systemzustands führt zur Produktion von *orders* und *conditions*. *Orders* werden an Module der Zwischenschicht und der reaktiven Ebene gesendet, um ihr Verhalten dem neuen Systemzustand anzupassen. *Conditions* werden an Module der deliberativen Ebene gesendet, um sie über die Zustandsänderung zu informieren.

Einen genauen Überblick der Zustände und Transitionen liefert die Darstellung 6.11. Die Farbe der Überschrift der jeweiligen Transition gibt die Quelle derselben an – so kann der ESV z.B. von dem *InteractionAttention*-Zustand in den *PersonAttention*-Zustand übergehen, wenn der Dialog ein Stopsignal liefert. Eine genauere Erklärung der Grafik würde den Rahmen dieser Arbeit sprengen, trotzdem möchte ich an dieser Stelle einen kurzen Überblick über die möglichen Zustände geben:

- **PersonAlertness:** BIRON hat in diesem Zustand keinen Kommunikationspartner. In diesem Fall übernimmt das PTA (*person tracking and attention system*) die Kontrolle, um potentielle Kommunikationspartner zu finden.

- **PersonAttention:** Wenn sich ein Kommunikationspartner bei BIRON „anmeldet“, also den Roboter begrüßt, richtet dieser seine volle Aufmerksamkeit auf die Person. In diesem Fall wird versucht, das Gesicht des Benutzers mit der

[3] BLOB ist eine Abkürzung für „*binary large object*“ und bezeichnet Binärdatenobjekte.
[4] *augmented finite state machine*

Pan-Tilt-Kamera zu verfolgen so wie ausschließlich gesprochene Sprache von der Person als für die Kommunikation relevant zu interpretieren.

- **PersonFollow:** BIRON folgt anhand des Laser-Distanzscanners dem Benutzer.
- **InteractionAttention:** BIRON wurde darauf aufmerksam gemacht, dass gleich ein Objekt gezeigt wird, und versucht Zeigegesten zu erkennen. Dieser Modus ist teilweise obsolet, da mittlerweile auch im *PersonAttention*-Zustand Objekte gezeigt werden können.
- **ObjectAttention:** BIRON versucht ein (gezeigtes) Objekt zu finden und die Ansicht dieses Objektes zu speichern.
- **GoTo:** In diesem Zustand übernimmt die Lokalisation die Navigation des Roboters zu einem vom Benutzer vorher spezifizierten Ort.

Der ESV stellt ein mächtiges Werkzeug zur Verwaltung der Systemzustände dar, arbeitet allerdings nahezu ausschließlich auf einer Metaebene – er dient im Wesentlichen der Verwaltung des Systems und nicht der Verarbeitung von sensorieller Information. Im Gegensatz dazu arbeiten die meisten anderen Module auf Sensordaten, die ggf. durch weitere Module vorverarbeitet wurden.

Person tracking and attention system Das *person tracking and attention system* (PTA) ((Lang, 2005), (Lang u. a., 2003), (Fritsch u. a., 2003) und (Kleinehagenbrock u. a., 2002)) umfasst eine Reihe von Modulen, die sich mit der Aufmerksamkeitssteuerung des Robotersystems befassen, womit im Speziellen die Aufmerksamkeit des Roboters gegenüber Menschen gemeint ist.

Wie geschildert soll der Roboter so lange potentielle Kommunikationspartner betrachten, bis ein solcher den Roboter anspricht und begrüßt, was das Signal für die begonnene Kommunikation darstellt. Von diesem Moment an bis zur Verabschiedung soll der Roboter nur diesen Kommunikationspartner selektieren.

Die Problematik bei der Erkennung von (potentiellen) Kommunikationspartnern ist, dass unter Verwendung heute üblicher Algorithmen keine einzelne Modalität zur Erkennung einer Person ausreicht. Aus diesem Grund verwendet das PTA drei verschiedene Sensoren (bzw. Perzepte), um Personen zu entdecken. Diese sind im Einzelnen:

- eine Gesichtserkennung mit Hilfe der Pan-Tilt-Kamera
- die Stereomikrophone (Schallquellenlokalisierung)
- der Laserscanner (Distanzmessung)

Zur **Wahrnehmung von Gesichtern** mit Hilfe der Pan-Tilt-Kamera wird ein Detektor auf Basis des Viola-Jones-Algorithmus (Viola und Jones, 2002) verwendet, wobei zur Konstruktion der Klassifikatoren AdaBoost (Freund und Schapire, 1997) benutzt wurde.

Die **Sprecherlokalisation** mit Hilfe der Stereomikrophone erfolgt auf der Basis einer Cross-Powerspektrum-Phasenanalyse (Giuliani u. a., 1994). Dieses Verfahren versucht anhand der unterschiedlichen Signallaufzeiten des Schalls zu den jeweiligen Mikrofonen die Schallquelle zu ermitteln.

Die **Personenlokalisation** anhand des Laserdistanzscanners ist darauf trainiert, das charakteristische Muster von Beinpaaren im Scanbereich zu erkennen.

Das PTA verwendet einen so genannten *anchoring*-Ansatz (Coradeschi und Saffiotti, 2001) (Lang u. a., 2003): Einer abstrakten Repräsentation von Objekten innerhalb der Welt wird ein mustererkennender Prozess zugeordnet, der versucht, Objekte aus Sensordaten zu ermitteln. Auf diese Weise wird zu jedem Zeitpunkt t ein 3-Tupel generiert, bestehend aus dem das Objekt repräsentierenden Symbol, einer Repräsentation der wahrnehmbaren Eigenschaften des Objektes und einem Perzept desselben. Wenn das Perzept des Objektes vorhanden ist, wird die Objektrepräsentation aktualisiert, ansonsten werden ggf. nur algorithmische Approximationen der vermuteten Objekteigenschaften gespeichert.

Das Besondere an dem auf BIRON realisierten *anchoring*-Ansatz ist die multimodale Verankerung von Objekten: Da ein Perzept meist nicht zur Wahrnehmung von Personen ausreicht, werden sie anhand dreier Modelle miteinander fusioniert:

- das **Kompositionsmodell** beschreibt die räumlichen Zusammenhänge der Komponenten in Bezug auf das zu perzipierende Objekt. So können in diesem Modell z.B. die maximalen Distanzen von den Beinpaarperzepten zum Gesichtsperzept festgelegt werden.
- das **Bewegungsmodell** beschreibt die Bewegungsarten des Objektes und erlaubt somit Rückschlüsse auf zukünftige Aufenthaltsorte.
- das **Fusionsmodell** gleicht die unterschiedlichen Arbeitsgeschwindigkeiten der einzelnen Perzepte aus, um so zeitgleiche Repräsentationen zu ermöglichen.

Die einzelnen Zustände der Aufmerksamkeitssteuerung werden in Analogie zu den Zuständen des ESV ebenfalls in einem Finiten Automaten gespeichert, der in der folgenden Grafik 6.3 dargestellt ist. Der Auslöser der Transition ist entweder der ESV (Grafik: ES) oder das PTA (Grafik: PT). An dieser Stelle möchte ich nur auf das Verhalten im Bottom-Up-Modus eingehen, in dem das Verhalten von BIRON im Wesentlichen durch das PTA bestimmt wird:

BIRON befindet sich standardmäßig im *sleeping*-Modus (keine Bewegung, Kamera gesenkt). Geräusche, die über eine einstellbare Zeit anhalten, lassen ihn in den *awake*-Modus übergehen. In diesem Modus richtet BIRON seine Aufmerksamkeit zwischen den von ihm wahrgenommenen Personen hin und her (Ausrichtung der Basis und Pan-Tilt-Kamera). Im Falle einer einzelnen, sprechenden Person wendet er sich dieser zu (*listening*), bis diese für länger als 2 Sekunden (konfigurierbar) nicht spricht, was ihn wieder in den *alert*-Modus versetzt. Wenn die sprechende Person BIRON begrüßt, wendet er seine volle Aufmerksamkeit der Person zu, bis er diese entweder nicht mehr lokalisieren kann oder sich die Person von ihm verabschiedet.

Für die Themenerkennung wichtig sind insbesondere die Zustände der Kommunikationspartner. So liefern die einzelnen Mustererkennungsalgorithmen z.B. die Information, ob BIRONs Gegenüber ihn anblickt, sich in Bewegung befindet oder spricht. Das *anchoring* ist besonders nützlich, da es erlaubt, Zustände einzelnen Individuen zuzuordnen – so würde BIRON nicht fehlerhafterweise annehmen, dass sein Kommunikationspartner redet, wenn stattdessen eine andere Person im Raum dies tut. Außerdem macht das PTA situierte Kommunikation erst in vieler Hinsicht überhaupt möglich, so dass diese Komponente eine elementare Voraussetzung für eine erfolgreiche Themenerkennung darstellt.

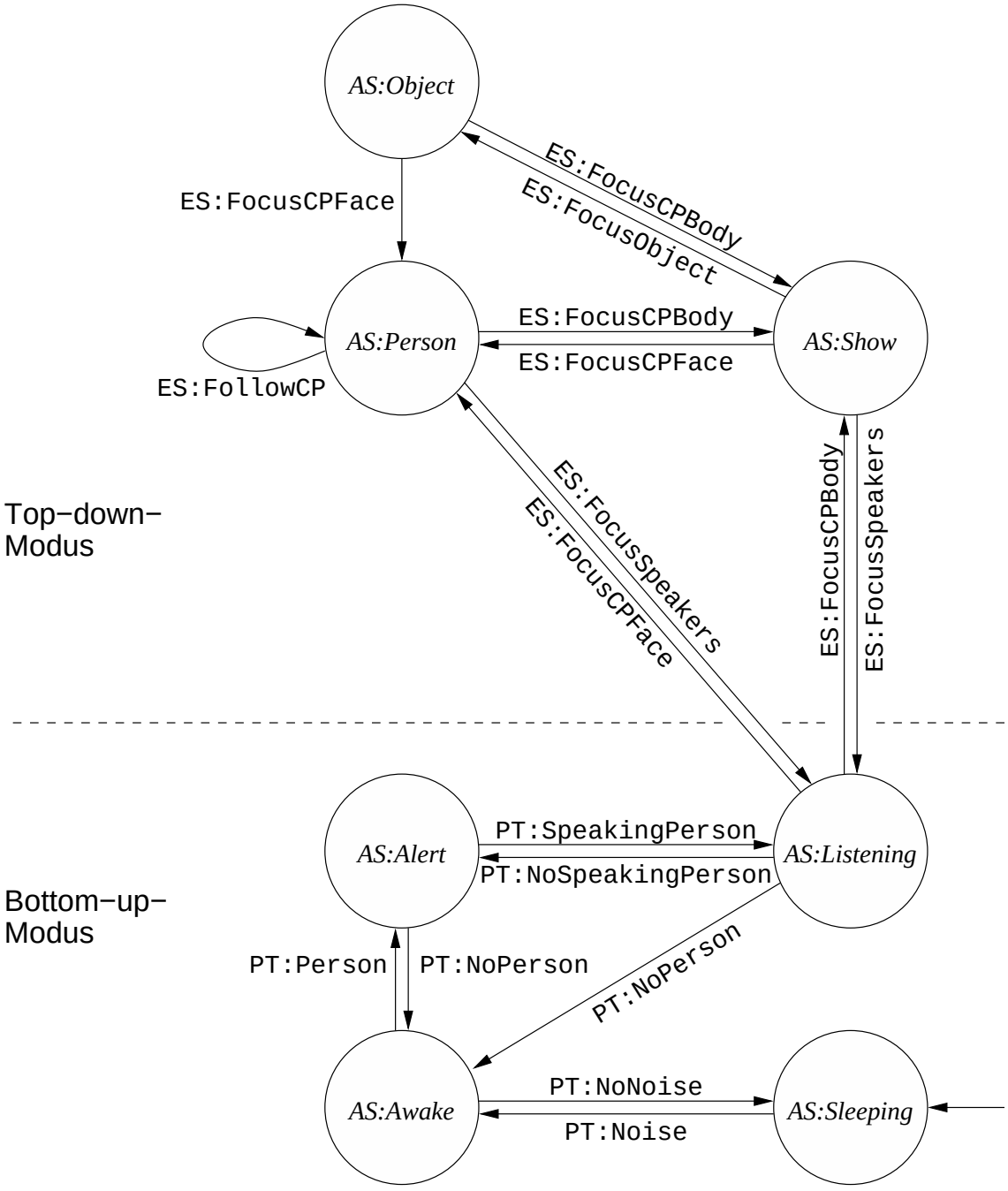

Abbildung 6.3: Aufmerksamkeitszustände (AS) und -transitionen von BIRON (PTA)

Speech Recognition, Speech Understanding und Dialogsystem Diese Modulgruppe umfasst die komplette Sprachein- und Ausgabe des Robotersystems. In Darstellung 6.2 nicht explizit erwähnt wurde z.B. das **Text-To-Speech-System**, welches direkt von dem Dialogmodul angesteuert wird. Es ermöglicht, in Strings kodierte Wörter direkt in Sprache umzusetzen. Dieses geschieht in zwei Schritten (Lömker, 2004, S.43ff): In einem ersten Schritt werden mit Hilfe der Phonemgenerierungsmethoden des Text-To-Speech-Systems Festival (Black u. a., 1999) die jeweiligen Silben in Phoneme umgesetzt. Für den zweiten Schritt – der Umwandlung der Phoneme in Laute – wird jedoch nicht Festival, sondern MBROLA (*Multi Band Resynthesis OverLap Add*) (Dutoit u. a., 1996) verwendet.

Die **Spracherkennung** (*speech recognition*) auf BIRON ist HMM-basiert (Fink, 1999). Sie wird zusätzlich von einem System zur Sprecherlokalisierung unterstützt (SPLOC – *speaker localisation*), welches anhand der Analyse der Differenzen des Eingangssignals der beiden Stereomikrophone die Schallquelle lokalisieren, und auf diese Weise gegebenenfalls Störschall besser ausfiltern kann (Hohenner, 2005) (Hohenner u. a., 2003).

Die Spracherkennung generiert anhand des Lexikons Worthypothesen, die dann an das **Sprachverstehensmodul** (ASU) (*speech understanding*) weitergeleitet werden. Diese Komponente (Hüwel u. a., 2006) (Hüwel und Wrede, 2006) dient der robusten semantischen Interpretation der – potentiell fehlerhaft erkannten – Sprache. Den Kern des Sprachverstehensmoduls bilden sog. SSUs (*situated semantic units*). SSUs repräsentieren basale semantische Konzepte wie „zeigen“ oder „Tisch“. Während des Sprachverstehensprozesses werden Worten der Benutzeräußerung SSUs zugeordnet. SSUs sind miteinander stark (*mandatory*) oder optional verbunden. Ein Beispiel wäre das Verb „zeigen“, welches u.a. stark mit einem zu zeigenden Objekt verknüpft ist. Wird in der Analyse der vorliegenden Äußerung keine passende Objekt-SSU gefunden, ist das Dialogsystem dazu angehalten, dieses in einer Rückfrage zu ermitteln[5].

SSUs bieten den Vorteil, in einer Ontologie strukturierbar zu sein – so kann z.B. die Objekt-SSU für „Tisch“ von einer allgemeinen Objekt-SSU abgeleitet werden und somit Eigenschaften aller Objekte erben.

Ein weiterer großer Vorteil liegt in der Unabhängigkeit von einer einzigen Modalität: So kann die SSU des Wortes „*this*“ optional mit einer Zeigegeste verbunden sein, so dass die semantische Information über die Gesten- und Objekterkennung ergänzt wird. Alternativ können auch nur gezeigte, nicht verbal repräsentierte Objekte durch eine SSU vertreten werden, um so die Anforderungen anderer SSUs zu erfüllen.

Die folgende Abbildung 6.4 (Hüwel und Wrede, 2006, S.4) zeigt die SSU für das Verb „*showing*“ (zeigen). „Zeigen“ ist notwendigerweise mit einer „*actor*“-SSU und einer „*object*“-SSU verbunden („Ich (*actor*) zeige den Tisch (*object*)“)[6]. Optional kann noch eine weitere involvierte Person („Ich zeige dir (*person_involved*) den Tisch“) und eine Zeit spezifiziert werden. Der Umstand, dass bestimmte SSUs wie z.B. die zweite involvierte Person nicht notwendigerweise spezifiziert werden müssen, obwohl diese üblicherweise mit angegeben werden, erlaubt die robuste Analyse von Äußerungen, deren Worte nur teilweise verstanden wurden. Das Fehlen obligatorischer SSUs kann durch Rückfragen korrigiert werden.

[5] In bestimmten Fällen werden auch in benachbarten Äußerungen passende SSUs gesucht.

[6] Das Beispiel ist der Einfachheit halber auf Deutsch, da sich die semantische/grammatische Struktur des Deutschen und des Englischen in diesem Fall sehr ähnlich sind.

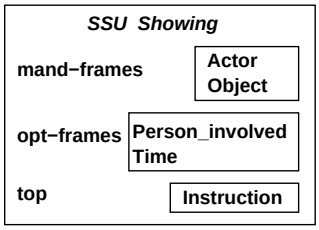

Abbildung 6.4: Schematische SSU „*showing*“ (zeigen)

```
<metaInfo>
<time>1125573609635</time>
<status>full</status>
</metaInfo>
<semanticInfo>
<u>what can you do</u>
<category>query</category>
<content>
  <unit = Question_action>
     <name>what</name>
     <unit = Action>
        <name>do</name>
        <unit = Ability>
           <name>can</name>
           <unit = Proxy>
              <name>you</name>
        ...
```

```
<u>this is a green cup</u>
<category>description</category>
<content>
  <unit = Existence>
     <name>is</name>
     <unit = Object_kitchen>
        <name>cup</name>
        <unit = Potential_gesture>
           <name>this</name>
        </unit>
        <unit = Color>
           <name>green</name>
        </unit>
       ...
```

Abbildung 6.5: Partielle ASU-Ergebnisse der Analyse der Sätze „*what can you do*“ und „*this is a green cup*“.

Die Generierung einer semantischen Repräsentation entsteht durch einen Unifikationsprozess, in dem sich das ASU für eine Repräsentation entscheidet. Im Fall von konkurrierenden Repräsentationen (Ambiguitäten) erfolgt dies anhand eines Punktevergabesystems, bei dem die Repräsentation mit den meisten Punkten als Sieger hervorgeht.

Für die Themensegmentierung besonders wichtig ist die Spezifikation von Oberkategorien in bestimmten SSUs (im Beispiel: *instruction*). Anhand dieser kann auf den Zweck der Äußerung im Dialog geschlossen werden. Zum Zeitpunkt der Erstellung dieser Arbeit existieren die Oberkategorien *confirmation, correction, description, instruction, negation, query* und *socialization*, obgleich weitere eingeführt werden können.

Die Ausgabe der ASU erfolgt als hierarchisch gegliederte XML-Struktur, die alle semantischen Informationen, Zeitstempel, involvierte Wörter etc. enthält. Ein Beispiel aus (Hüwel und Wrede, 2006, S.6) ist in Abbildung 6.5 ersichtlich.

Auch wenn im Rahmen dieser Arbeit nicht detailliert auf dieses Beispiel eingegangen werden kann, ist trotzdem klar erkennbar, wie sich die einzelnen Worte (*name*)

und SSU-Typen (*unit*) in die semantische Analyse eingliedern. Auf der linken Seite sind Metainformationen wie der Startzeitpunkt der Äußerung erkennbar; die *category* ist identisch mit der übergeordneten Kategorie der unifizierten SSUs.

Dem **Dialogsystem** (Li u. a., 2006) (Li, 2006) liegt die so genannte *common ground*-Theorie (Clark, 1992) zugrunde. Laut dieser Theorie müssen während einer Kommunikation die kommunizierenden Agenten ihre mentalen Zustände, also ihre Einstellungen bezüglich ihrer Ziele, Intentionen und Aufgaben koordinieren (*grounding*). Um dies zu erreichen werden *contributions* (Beiträge) formuliert, die stets in zwei Teile untergliedert sind: *presentation* und *acceptance*. Die *acceptance* ist die Bestätigung des Verständnisses der *presentation* und wird daher stets von dem Kommunikationspartner geäußert, der nicht die *presentation* formuliert hat.

Presentation und *acceptance* können im einfachsten Fall durch einzelne Äußerungen der Kommunikationspartner realisiert sein:

Beispiel 6.1
Presentation: *„dies ist ein Schrank“*
Acceptance: *„OK“*

Allerdings können sie auch durch Gesten oder Mimik geäußert werden, außerdem kann eine erneute *presentation* eine eigene *acceptance* darstellen:

Beispiel 6.2
Presentation: *„ich bin Jan“*
Acceptance: *„und ich heiße Anna“*

Bei Missverständnissen kann die *acceptance* erst nach verschiedenen Rückfragen, die jeweils selbst wieder eigene *presentation*/*acceptance*-Paare darstellen, erfolgen:

Beispiel 6.3
Presentation 1: *„ich bin Jan“*
Presentation 2: *„wie bitte?“*
Acceptance 2: *„ich heiße Jan“*
Acceptance 1: *„OK“*

Ein *presentation*/*acceptance*-Paar wird nach (Cahn und Brennan, 1999) als *exchange* (Austausch) bezeichnet. Somit stellen Austausche die basalen Einheiten eines *grounding*-Prozesses dar.

Ein *exchange* (Austausch) gilt als *grounded*, wenn er vollständig ist. Die Austausche, die diese Bedingung nicht erfüllen, werden in einem Stack abgelegt. Eine der Hauptaufgaben des Dialogsystems besteht darin, Benutzeräußerungen als *presentation* oder *acceptance* einzuordnen. Dazu ist der Stackspeicher nützlich, da er die zu erwartenden Dialogelemente und Beziehungen zwischen den einzelnen Austauschen speichert. In dem Dialogsystem existieren vier verschiedene Beziehungen zwischen Austauschen:

- **Default:** Der neue Austausch ist unabhängig von dem vorigen.
- **Support:** Ein neuer Austausch wird initiiert, um den vorigen beenden zu können (z.B. bei Nachfragen).
- **Correct:** Im Fall von Missverständnissen kann ein Austausch den vorigen korrigieren.

- **Delete:** Aufgrund von Kommunikationsschwierigkeiten o.ä. wird der letzte Austausch abgebrochen, ein *grounding* findet nicht statt.

Ein Austausch, der *grounded* wurde, gilt als beendet und wird sofort vom Stack entfernt. Ein völlig leerer Stack kennzeichnet die Grenze eines Diskurssegments.

Wissen über die Relationen zwischen den Austauschen können in begrenztem Maße zur Segmentierung von Diskursen nach Themen eingesetzt werden, da z.B. eine Support-Relation einen Themenwechsel verbietet.

Eine weitere wichtige Leistung des Dialogsystems von BIRON besteht in der Auflösung multimodaler Objektreferenzen mit Hilfe der Objekt- und Gestenerkennung (Li u. a., 2005). Verbale Auslöser wie „*this*“ können eine Objektsuche initiieren, die dann zu einer Verknüpfung der Benutzeräußerung mit der Objekt-ID führt.

Object learning und gesture detection BIRON verfügt über eine ansichtsbasierte **Objekterkennung** (*object learning*) (Haasch u. a., 2005), die in Kombination mit einer trajektorienbasierten **Gestenerkennung** (*gesture detection*) (Hofemann u. a., 2004) für Zeigegesten arbeitet. Während einer Zeigegeste wird die Hand des Benutzers anhand der Hautfarbe verfolgt und ein Bereich berechnet, in dem sich das gezeigte Objekt mit großer Wahrscheinlichkeit befindet. Anschließend kann das *object learning* anhand der vom Benutzer spezifizierten Farbe des Objektes das bezeichnete Objekt innerhalb des Bereichs finden.

Problematischerweise ist somit zwar die Grundlage für die Wiedererkennung von Objekten gegeben, selbige wurde aber noch nicht in BIRON implementiert. Aus diesem Grund wurde in den in Abschnitt 6.6 beschriebenen Experimenten die Objektwiedererkennung simuliert (siehe dort).

Localisation, hardware control, robot data server Diese Module dienen unterschiedlichen Zwecken, stehen aber nur in sehr geringem Zusammenhang mit dem Themenerkennungssystem, weswegen ich an dieser Stelle ihre Funktion nur kurz umreißen werde.

Die **localisation** erlaubt die odometrie- und ansichtsbasierte Navigation zu Orten, die BIRON mit Hilfe des Kommunikationspartners kennen gelernt hat (Spexard u. a., 2006).

Die **hardware control** (HWC) dient der Ansteuerung der Pan-Tilt-Kamera, so wie der Roboterbasis.

Der **robot data server** (RDS) liefert Informationen über den Zustand der Kameras (Zoom, Ausrichtung, etc.) und des Laserscanners (Daten).

6.2 Eingliederung des Themenerkennungssystems in die Systemarchitektur

Im Rahmen der ersten Implementierung des Themenerkennungssystems in die Architektur von BIRON stellte sich die Frage, wie das Themenerkennungssystem zur Erfüllung seiner Aufgaben Nutzen aus dem bestehenden System ziehen konnte.

Das Themenerkennungssystem benötigt für zwei der von ihm zu bewältigenden Aufgaben Informationen aus dem Gesamtsystem. Diese Aufgaben sind im Einzelnen:

- Sammeln der multimodalen Symbole, die als Themenindikatoren eingesetzt werden

- Segmentierung der produzierten Symbole in Gruppen, die (wahrscheinlich) ein Thema besitzen.

Der erste Punkt bezeichnet im Wesentlichen die Anbindung an die Ergebnisse der Vorverarbeitungsstufen, die diese Symbole wahrnehmen bzw. generieren. Der zweite Schritt kann auf vielfältige Weise geschehen. Im Kontext des entwickelten Systems wurde er dadurch realisiert, dass die Dialoge zwischen Mensch und Roboter in thematische Zeitabschnitte unterteilt werden, denen das System dann weitere Symbole zuordnet, bzw. Wörter durch andere Symbole ersetzt (z.B. Objektreferenzen).

Während in Abschnitt 6.3 detailliert auf das Problem der Segmentierung und die im Prototypen realisierte Lösung eingegangen wird, möchte ich an dieser Stelle kurz die Lösung der ersten Aufgabe – der Sammlung von multimodalen Symbolen, die zur Themenerkennung verwendet werden können – eingehen.

Genau wie in den *offline*-Experimenten besteht die Einbindung der Multimodalität in einer Auflösung von Objektreferenzen, die verbal durch den Benutzer unternommen werden. In der momentanen Architektur von BIRON ist es notwendig, dass eine Objektbezeichnung mit einer Zeigegeste und einer Farbbezeichnung einhergeht (*„this is* (Geste) *a green pen“*). Fehlt letztere, wird der Benutzer in einer Rückfrage aufgefordert, diese zu nennen.

Problematischerweise ist auch noch keine ansichtsbasierte Objekt-Wiedererkennung integriert, die für eine Unterstützung der Themenerkennung durch Objektreferenzen nötig ist. Anaphorische Rückbezüge auf bekannte Objekte können in geringem Maße von dem Dialogsystem durchgeführt werden, so dass z.B. Personalpronomina durch ihre Objektreferenz-IDs ersetzt werden können.

Aus diesen Gründen wurden für das Themenerkennungssystem zwei Betriebsmodalitäten vorgesehen: In einem Modus werden ausschließlich Worte verarbeitet, es handelt sich also um eine unimodale Verarbeitung. Um zukünftigen Entwicklungen Raum zu bieten kann jedoch die Verarbeitung von Objektreferenzen – die in der Kommunikation zwischen den Modulen schon vorgesehen ist – aktiviert werden. In diesem Fall werden innerhalb des Themenerkennungsmoduls die Objektreferenzen in Form von IDs in die jeweiligen Äußerungen aufgenommen, die Löschung der referenzierenden Wortsymbole erfolgt in einem zweiten, unabhängigen Schritt.

Da die Implementierung einer Objekt-Wiedererkennung ein lösbares technisches Problem darstellt, welches aber auf BIRON noch nicht umgesetzt wurde, wurden die in diesem Kapitel beschriebenen Experimente im unimodalen Modus durchgeführt. Allerdings wurde die Objekterkennung durch einen *Wizard-of-Oz* simuliert, so dass *de facto* ein multimodaler Betrieb getestet wurde. Eine genaue Beschreibung dieses Vorgangs findet sich bei der Beschreibung der Experimente in Abschnitt 6.6.

Die zur Themenerkennung herangezogenen Worte wurden einfach über das Dialogsystem dem Spracherkennungssystem entnommen, wobei allerdings nur Benutzeräußerungen und keine Roboteräußerungen verwendet wurden. Dies geschah nicht zuletzt aus dem Grund, dass in dem momentanen System keine roboterinitiierten Themenwechsel durchgeführt werden können. Weiterhin werden vom Sprachverstehen oder vom Dialog als fehlerhaft markierte Benutzeräußerungen nicht verarbeitet.

Abbildung 6.6 zeigt die Verortung des DTT-Moduls in das Gesamtsystem. Die

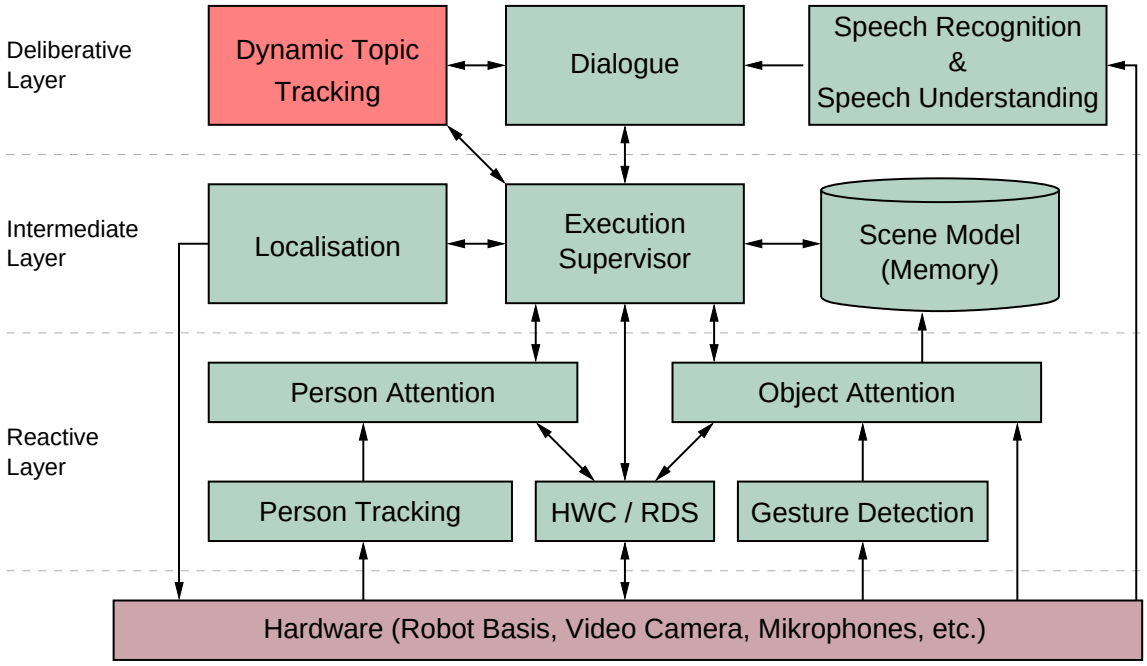

Abbildung 6.6: Neues Drei-Schichten-Modell der Software von BIRON

Platzierung erfolgt aufgrund des Inferenzen bildenden Charakters des Moduls auf der deliberativen Ebene, von der aus es in direktem Kontakt zu dem Dialogsystem und dem ESV steht. Der Datenaustausch mit den anderen Modulen erfolgt über dieselben. Diese Form der Anbindung erlaubt es, dass das Dialogmodul direkt von der Ausgabe des DTT-Moduls profitiert, aber zumindest prinzipiell auch eine Kommunikation vom DTT zum ESV möglich ist. Diese beschränkt sich z.Zt. auf ein Lebenszeichen, kann aber in Zukunft dazu verwendet werden, beliebige Module über das aktuelle Thema zu informieren.

6.3 Segmentierung

Im vorliegenden Abschnitt möchte ich auf die im Prototypensystem realisierte Segmentierung eingehen.

6.3.1 Vorarbeiten zur automatischen Segmentierung

Wie in Abschnitt 3.1.2 auf Seite 26ff geschildert wurde, existieren verschiedene Verfahren zur Segmentierung von Texten bzw. Sprachströmen. Generell unterschieden wurde zwischen textkohäsionsbasierten Verfahren – also Verfahren, die anhand syntaktischer bzw. semantischer Eigenschaften von Texten Textkohäsion messen, um daraus Rückschlüsse auf Themenwechsel bzw. Grenzen von Themenabschnitten zu schließen – und Multi-Quellen-Verfahren. Diese *catch all*-Klasse umfasst ein stetig wachsendes Spektrum von verschiedenen Analysemethoden von Text- und Audioquellen.

Kohäsionsbasierte Verfahren Prinzipiell lassen sich Standardverfahren zur Textkohäsionsmessung auch auf einem multimodal kommunizierenden Robotersystem einsetzen, nachdem eine Spracherkennung die gesprochene Sprache in Text umgewandelt hat. Potentiell problematisch sind dabei neben Spracherkennungsfehlern insbesondere zwei Punkte:

1. Die Grundstruktur des zu segmentierenden Textes ist der Dialog, während viele Textsegmentierungsverfahren für aufsatzartige Texte (Zeitungsartikel o.ä.) entwickelt wurden.

2. Es liegt immer nur ein bestimmter Teil des Textes vor, wobei die zu findenden Grenzen nur an dessen Ende liegen können.

Es wurden zwei Verfahren zur Textsegmentierung angesprochen: Blockvergleich und Vokabeleinführungsanalyse. Beide Verfahren arbeiten jedoch mit Fenstern, die über einen Text geschoben werden und Themenwechsel in ihrer Mitte erkennen sollen. Das Fehlen großer Teile zukünftiger Information (also der „rechten“ Seite des Fensters) macht die Anwendbarkeit dieser Verfahren prinzipiell schwierig.

Trotzdem wäre es möglich, die Segmentierung durch Hinweise wie wiederkehrendes Vokabular zu unterstützen; eine geringe Menge an Text, der nach einem möglichen Themenwechsel geäußert wurde, steht meistens zur Verfügung. Auch könnte eine Analyse anaphorischer Beziehungen hilfreich sein – so ist eine anaphorische Referenz (z.B. durch Personalpronomina) über Themengrenzen selten und ihre Anwesenheit kann als Kriterium für die Verschmelzung zweier Äußerungen zu einem Kommunikationssegment dienen.

Im Rahmen dieser Arbeit wurde nicht mit vergleichbaren Verfahren experimentiert; ich gehe aber davon aus, dass eine Analyse anaphorischer Beziehungen von Nutzen wäre. Letztere wurde aber noch nicht in das Dialogsystem integriert.

Multi-Quellen-Verfahren Auch die aus verschiedenen Eingabequellen schöpfenden Verfahren, die in der Vergangenheit entwickelt wurden, sind nicht 1:1 auf die in HRI vorherrschende Situation anwendbar. Prosodiebasierte Verfahren liefern bei frei gesprochener Sprache schlechtere Ergebnisse als z.B. pausenbasierte (Shriberg u. a., 2000), pausenbasierte lassen sich aber aufgrund der dialogischen Struktur von Mensch-Roboter-Interaktionen nicht anwenden. Aus diesem Grund wurde eigens für diese Arbeit und den Roboter BIRON ein Multi-Quellen-Verfahren entwickelt und implementiert. Der Grundgedanke bei der Entwicklung dieses Systems bestand insbesondere darin, ein robustes System zu entwickeln, welches aufgrund der nicht vorhandenen Trainingsdaten ohne vorheriges Lernen auskommt. Ich möchte es im Folgenden darstellen.

6.3.2 Realisierung der Segmentierung auf BIRON

In dem im Rahmen dieser Arbeit entwickelten Prototypen einer *online*-Themenerkennung auf dem Roboter BIRON wurden Informationen verschiedener Module des Robotersystems verarbeitet, um erste Schritte in Richtung einer optimalen automatischen Segmentierung zu unternehmen. Die verwendeten Informationsquellen waren im Einzelnen:

- **Attention:** Unterbrechungen in der Aufmerksamkeit des Benutzers deuten auf Themenwechsel hin, da sie sowohl Störungen durch Ereignisse außerhalb der Kommunikation oder aber auch interne Themenwechsel – z.B. durch einen Ortswechsel – anzeigen. Wichtig ist an dieser Stelle auch die Information, ob eine Periode geteilter Aufmerksamkeit (*joint attention*) vorliegt, innerhalb derer fast nie ein Themenwechsel vorliegt.

- **Dialogsystem:** Das Dialogsystem liefert potentiell nützliche Informationen über Themenwechsel anhand der Dialogstruktur.

- **Sprachverstehen (ASU):** Bestimmte Äußerungen des Benutzers lassen auf einen Themenwechsel schließen (z.B. im einfachsten Fall: „lass uns das Thema wechseln“).

Jeder zu einem Zeitpunkt von einem Modul festgestellte Themenwechsel wird in eine als Java-*TreeMap* realisierte Liste eingetragen, anhand derer das Trackingmodul mögliche Themenwechsel erkennt. Themenwechsel können dabei wie im BITT-Korpus immer nur zwischen (und nicht innerhalb von) Äußerungen stattfinden. Äußerungsgrenzen werden wie geschildert von dem VAD-System erkannt.

Themenwechselindikation durch das attention-System und den ESV

Die Informationen des *attention*-Systems und des ESVs wurden in Kombination benötigt, um eine Themensegmentierung anhand der Aufmerksamkeit von Kommunikationspartnern zu ermöglichen. Während der Erstellung des BITT-Korpus fiel auf, dass Situationen, in denen die Versuchsperson von BIRON wegsah, häufig einen Themenwechsel kennzeichneten. Das *attention*-System gibt anhand einer Analyse der von der Pan-Tilt-Kamera gelieferten Bilder Aufschluss darüber, ob der Kommunikationspartner BIRON ansieht oder nicht.

Problematisch sind allerdings Fälle von *joint attention*, in denen die Aufmerksamkeit beider Kommunikationspartner auf ein Objekt gerichtet ist. Im Rahmen des *home tour*-Szenarios und BIRONs kommunikativen Fähigkeiten ist dies der Fall, wenn der Benutzer BIRON ein Objekt zeigt. In diesem Fall wechselt der interne Zustand von BIRON in den *object attention*-Modus; BIRON richtet in diesem Modus seine Pan-Tilt-Kamera auf das gezeigte Objekt und ist nicht mehr in der Lage, die Blickrichtung des Kommunikationspartners wahrzunehmen.

Die im Prototyp zur Anwendung gekommene Segmentierung sieht vor, dass alle Zeitperioden, in denen der Kommunikationspartner BIRON nicht ansieht, als Themengrenze angesehen werden, sofern das System sich nicht im *object attention*-Modus befindet. Der Anfang des Zeitperiode gilt in diesem Fall als Zeitpunkt des Themenwechsels.

Prinzipiell lassen sich aus dem *attention*-System noch weitere Informationen über den Kommunikationspartner gewinnen, die auf einen Themenwechsel hindeuten. Ein Beispiel dafür wäre der Bewegungszustand: Fortbewegung auf Seiten des Benutzers kann in *home tour*-Szenarien einen Themenwechsel kennzeichnen. Aufgrund möglicher Fehlklassifikationen wird dieser Fall jedoch über das Sprachverstehen behandelt: Es werden also nicht alle Fälle, in denen sich der Kommunikationspartner bewegt, als Themenwechsel angesehen, sondern nur solche, in denen der Kommunikationspartner BIRON explizit auffordert, ihm zu folgen.

Der ESV dient noch in einer weiteren Hinsicht der Segmentierung: Situationen, in denen sich kein mit BIRON in Kommunikation befindliches Gegenüber findet – gekennzeichnet durch die Modi *sleeping* und *person alertness* – gelten in allen Fällen als Themenwechsel.

Themenwechselindikation durch das Dialogsystem

Die von dem Dialogsystem generierten Informationen werden im Rahmen des Prototypensystems nicht direkt zur Erkennung von Themenwechseln genutzt, es wird nur an jedem Ende eines Diskurssegments eine Aktualisierung der zum Training des Themenerkennungssystems angelegten Dialogsegmentdatenbank ausgelöst. Dies hat jedoch rein technische Gründe, da die Reihenfolge der Benutzeräußerungen aufgrund der Konstruktion des Dialogsystems nur am Ende eines Diskurssegments exakt bestimmt werden kann. Eine weiterführende Einbindung des Dialogsystems wäre möglich[7], würde aber idealerweise mit einer stärkeren Ausrichtung desselben auf eine thematische Orientierung des Dialogs einhergehen.

Für zukünftige Anwendungen jedoch wäre aus dem Dialogsystem vor allem die Information nützlich, ob ein Roboter-initiierter Themenwechsel vorliegt. Mögliche zukünftige Themenerkennungssysteme werden aus diesem Grund mehr auf das Dialogsystem als Informationsquelle für die Segmentierung zugreifen, als es der vorliegende Prototyp tut.

Der Hauptnutzen des Dialogsystems für die Themenerkennung liegt in der Fähigkeit desselben, verbale und gestische Objektreferenzen zu verschmelzen und so dem Themenerkennungssystem die Objektreferenzen zur Verfügung zu stellen.

Themenwechselindikation durch Sprachverstehen (ASU) und Schlüsselwörter, explizite Themenwechsel

Die aus dem Sprachverstehen gewonnenen Informationen stellten sich in mehrfacher Hinsicht als nützlich heraus. Zum einen können explizite Themenwechsel – die im *online*-System möglich sind – durch das Sprachverstehen erkannt werden. Zum anderen lässt die Abfolge bestimmter Kategorien[8] von Benutzer-Äußerungen auf Themenwechsel schließen.

Im Kontext des Prototypen wurde für jede Abfolge von zwei aufeinanderfolgenden Benutzeräußerungstypen definiert, ob ein Themenwechsel vorliegt oder nicht. Dies geschieht über eine frei konfigurierbare Matrix. Im Kontext des Prototypen wurden vor allem Äußerungen des Typs *socialisation* und des Typs *instruction* oftmals als themenwechselnd interpretiert, da diese im Kontext von BIRON Begrüßungen und Aufforderungen zu folgen beinhalten. Da in manchen Fällen *socialisation* auch ein „*thank you*“ bedeuten kann, wird dieser Prozess zusätzlich durch Schlüsselwortanalyse unterstützt. So wird eine *instruction*, die z.B. ein „*follow*“ enthält, im Rahmen des Prototyps als Themenwechsel interpretiert, da sie einen Ortswechsel (und damit wahrscheinlich einen Themenwechsel) nach sich zieht. Verschiedene andere *instructions* wie z.B. „*look here*“ werden nicht als themenwechselnd angesehen.

[7] vgl. Abschnitt 6.1.4 auf Seite 134ff

[8] Hiermit sind die weiter oben beschriebenen Oberkategorien von unifizierten SSUs gemeint.

Zusammenfassung und Auswirkungen erkannter Themenwechsel

Die folgende Tabelle fasst noch einmal alle Themenwechselauslöser auf, die im Prototyp zur Anwendung kommen:

- fehlender Blickkontakt bei nicht vorliegender *joint attention*
- Aufforderung, zu folgen
- Begrüßung, Abschied
- Fehlen eines Kommunikationspartners
- explizite (verbale) Themenwechsel

Die vorliegende Heuristik kann nur als Versuch gewertet werden, sich einem optimalen automatischen Segmentierungsprozess zu nähern. Wie ein solcher Prozess aussehen wird, ist zweifellos von den Fähigkeiten zukünftiger Robotersysteme und ihrer Einsatzgebiete abhängig.

Nach der Darstellung der Erkennung von Themenwechseln im Prototypensystem möchte ich noch kurz auf die Auswirkungen eingehen, die ein erkannter Themenwechsel auf das System hat. Im Wesentlichen führt ein erkannter Themenwechsel zu zwei Veränderungen im System:

1. Die für das Training des semantischen Raumes herangezogenen Kommunikationssegmente bestehen aus vom Kommunikationspartner geäußerten *interaction units*. Kommunikationssegmente werden aus den Benutzeräußerungen durch die erkannten Themenwechsel gebildet.

2. Jegliche *history*-Information wird gelöscht. Dies beinhaltet sowohl die Benennung des letzten erkannten Themas als auch die Gewichtung aktueller Klassifikationsvorgänge mit den Ergebnissen der vorangegangenen Klassifikation. Fokus und aktueller Themenname – die in Abschnitt 6.5 besprochen werden – werden auf diese Weise ebenfalls gelöscht.

Somit beeinflussen erkannte Themenwechsel nicht nur das Training des Systems, sondern auch direkt den Klassifikationsprozess.

6.4 Modifizierte Struktur des Themenerkennungssystems

Die Implementierung des Themenerkennungssystems in die Architektur von BIRON machte nicht nur eine Revision der *interfaces* zur Kommunikation mit der Softwareumgebung notwendig, sondern auch eine interne Umstrukturierung des Themenerkennungssystems. Während in den *offline*-Experimenten größere Mengen an Trainingsdaten auf einmal eingelesen und verarbeitet werden konnten, um anschließend zum Tracking einer zweiten Datenmenge herangezogen zu werden, verlangt ein implementiertes System nach einer permanenten Aktualisierung der Themeninformation. Da die Themenerkennung *online* Ergebnisse liefern muss, bestand die Notwendigkeit einer Auftrennung der einzelnen Prozesse in *threads*.

Die Grafik 6.7 verdeutlicht den Aufbau des *online*-fähigen Systems. Zur Kommunikation mit den anderen Komponenten wurde XCF herangezogen, welches einen *stack*-basierten Zugriff auf Systeminformationen ermöglicht.

In der Grafik ist jedes *thread* durch ein farbiges Kästchen mit spitzen Ecken gekennzeichnet. Kreise und Ellipsen kennzeichnen Module außerhalb des Themenerkennungssystems, während Kästchen mit abgerundeten Ecken Objekte darstellen, welche von den einzelnen *threads* aufgerufen werden können, um so z.B. global Daten abzulegen. Die jeweiligen Farben dienen nur der groben Gruppierung; die Module zur Verwaltung des Themenerkennungssystems sind gelb, die Kernmodule der Themenerkennung blau und die Module, die direkt mit den jeweiligen Informationsquellen in Verbindung stehen, grün markiert.

Die Pfeile kennzeichnen den Datenfluss, wobei keine Steuerungsnachrichten in die Grafik mit aufgenommen wurden. Die Beschriftungen der Pfeile geben Auskunft über die transportierten Daten. Zu beachten ist, dass rote Beschriftungen die Übermittlung des aktuellen Zustands oder einer neuen Information kennzeichnen, während schwarze Beschriftungen darauf hindeuten, dass an dieser Stelle auch ältere Informationen abgerufen werden können. So speichert die Datenbank („*Database*“) die aktuell (rot) eingehenden Segmente, um sie dann in ihrer Gesamtheit (schwarz) an die Berechnung des semantischen Raumes weiterzugeben.

Die folgende Aufzählung gibt einen kurzen Überblick über die einzelnen Informationstypen:

- **Äußerungen:** Hierbei handelt es sich um die (potentiell aus verschiedenen Modalitäten stammenden) Symbole, die während einer vom Spracherkenner erkannten (und durch die VAD segmentierten) Benutzer-Äußerung produziert wurden. Je nach Verortung im System können diese Symbole modifiziert worden sein (vgl. SymbolTransformer).

- **Segmente:** Segmente sind aufeinanderfolgende Äußerungen, die zu Themenabschnitten fusioniert wurden. Die Segmentierung erfolgt in den jeweiligen Modulen. Für die Symbole eines Segments gilt das gleiche wie für die einer Äußerung.

- **ETNamen:** Vom Benutzer explizit benannte Themennamen. Vgl. dazu Abschnitt 6.5.

- **TNamen:** Vom Themenerkennungssystem implizit bestimmte Themennamen. Vgl. dazu Abschnitt 6.5.

- **TWechsel:** Vom Themenerkennungsmodul erkannte Zeitpunkte von Themenwechseln.

- **Themenmodelle:** Wie im *offline*-System; Themenmodelle sind Cluster von Symbolvektoren samt den Distanzen aller bekannten Symbole zu dem Clustermittelpunkt.

- **Themen:** Entsprechen Themenmodellen, allerdings in Verbindung mit der Information, zu welchem Zeitpunkt das jeweilige Themenmodell das aktuelle Thema widerspiegelt.

- **Rest:** Die unerklärten Begriffe sollten selbsterklärend sein.

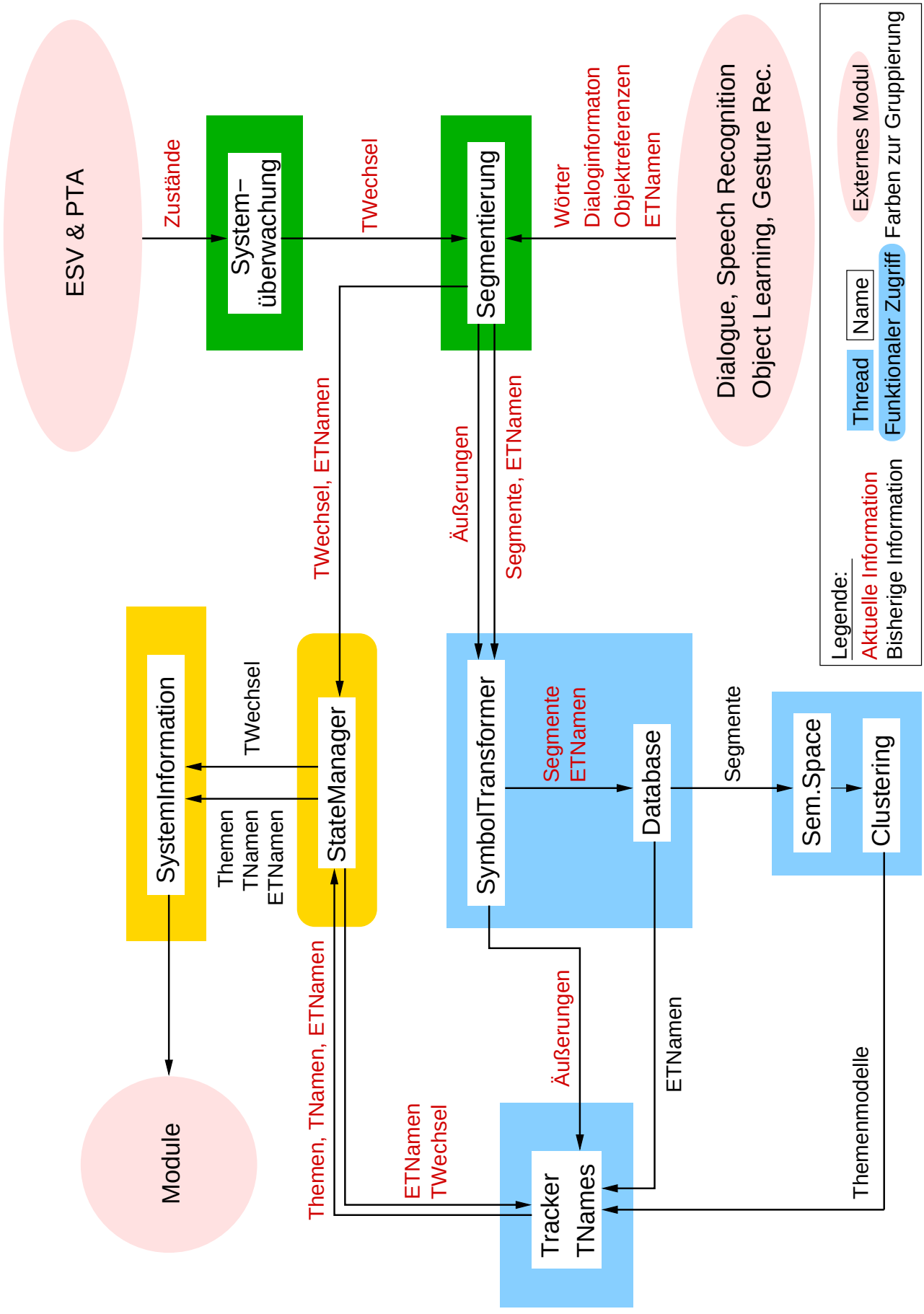

Abbildung 6.7: Schematische Darstellung der Themenerkennung

Die einzelnen Module erledigen dabei die folgenden Aufgaben:

- **Systemüberwachung:** Dieses Modul speichert die Zustände, die der ESV und das PTA einnehmen. Themenwechsel werden wie beschrieben durch die Abwesenheit von einem Kommunikationspartner gekennzeichnet oder durch Phasen, in denen der Kommunikationspartner über eine definierbare Zeit (default: 3 Sekunden) BIRON nicht ansieht. In Situationen, in denen BIRON auf ein Objekt blickt und somit die Blickrichtung des Partners nicht bestimmen kann, wird dieses Merkmal nicht beachtet. Gleiches gilt für Situationen, in denen der Kommunikationspartner schon auf ein Objekt blickt, um es anschließend zu zeigen (*joint attention*). Letztere Situationen gehen den erstgenannten häufig voraus. Um möglichst effizient auf die Änderungen der Systemzustände reagieren zu können, wurde die Systemüberwachung in einem eigenen *thread* realisiert.

- **Segmentierung:** Dieses Modul hat eine Vielzahl von Aufgaben. Anhand den von der Systemüberwachung erkannten Themenwechseln und den Themenwechselinformationen aus dem Dialogsystem ermittelt es die Zeitpunkte, an denen Themenwechsel stattgefunden haben. Mittels dieser Zeitpunkte werden alle innerhalb eines Zeitintervalls liegenden Benutzeräußerungen zu einem Segment zusammengefügt, wobei ggf. Objektreferenzen als multimodale Hinweise mit in die Äußerungen – und damit in die Segmente – aufgenommen werden. Die Segmentierung übermittelt neu erhaltene Äußerungen (zum Tracking) sofort, aus diesen Äußerungen zusammengefügte Kommunikationssegmente (zum Training) aber erst nach einem Themenwechsel weiter. Weiterhin extrahiert sie explizite Themennamen und markiert diese an den jeweiligen Segmenten. Die Segmentierung erfolgt in einem eigenen *thread*, da sie kontinuierlich über den Dialogeingabekanal wacht.

- **SymbolTransformer:** Dieses Modul übernimmt sowohl die Lemmatisierung von Worten als auch die Löschung von Funktionswörtern anhand einer Stopliste. Beides geschieht z.Zt. anhand von einfachen *lookup*-Tabellen, da BIRONs Vokabular verhältnismäßig klein ist. Dies ist jedoch keine Notwendigkeit; intelligentere Algorithmen können in das Modul integriert werden. Der SymbolTransformer leitet die Äußerungen an den Tracker, die Segmente mit den expliziten Themennamen an die Datenbank weiter.

- **Database:** Die Datenbank speichert alle eingehenden Segmente samt möglichen expliziten Themennamen, um sie der Berechnung des semantischen Raumes zur Verfügung zu stellen. Der SymbolTransformer und die Datenbank haben ein eigenes *thread*, da es wichtig war, sie von der Berechnung des semantischen Raumes zu trennen.

- **Sem.Space, Clustering:** Hierbei handelt es sich um die bekannte Berechnung des semantischen Raumes mit anschließender Bildung von Themenmodellen. Diese wird regelmäßig angestoßen, wenn die Datenbank um eine definierbare Anzahl neuer Segmente (Standard: eins) erweitert wurde. Nach Terminierung der Berechnung wird der Tracker benachrichtigt, so dass er sich aktualisieren kann. Die beiden Module laufen in einem eigenen *thread*, da ansonsten die potentiell lang dauernde Berechnung der Themenmodelle das restliche System blockieren würde.

- **Tracker und TNames:** Diese beiden Module tracken anhand der aktuellsten Themenmodelle die neu eingehenden Äußerungen und legen sowohl das Thema, also auch den potentiellen impliziten wie expliziten Themennamen fest. Nach der Aktualisierung der Themenmodelle werden auch Teile der Datenbank ausgelesen, um potentielle explizite Themennamen mit Themen in Bezug zu setzen, so wie implizite Themennamen für jedes Thema zu berechnen. Die Information über einen aktuellen, expliziten Themennamen so wie über aktuell erfolgte Themenwechsel, die zu einer Löschung der *history* führen, werden dem StateManager entnommen. Der Tracker ist eine der zeitkritischsten Komponenten der Themenerkennung, weswegen er zusammen mit der Namensbestimmung in einem eigenen *thread* läuft.

- **StateManager:** Dieses Modul ist als Objekt realisiert, auf das synchronisiert verschiedene *threads* zugreifen können. In diesem Objekt werden Informationen über alle bisherigen Themenwechsel so wie die aktuellen expliziten Themennamen gespeichert. Weiterhin wird das aktuelle Thema in diesem Objekt samt Themenname abgelegt.

- **SystemInformation:** Hierbei handelt es sich um ein generisches *interface* zur Kommunikation mit dem Gesamtsystem. In der Prototypenimplementierung konnte dieses Modul vom Dialogsystem dazu aufgefordert werden, Antworten auf Fragen nach dem aktuellen Thema zu generieren. Auch dieses Modul ist als *thread* realisiert, was aus technischen Gründen (Anbindung an XCF) erfolgte.

Wichtig ist in dieser Struktur vor allem der Umstand, dass die Trainingsprozesse (Sem. Space, Clustering) unabhängig von dem eigentlichen Trackingvorgang (Tracker) ablaufen. Erst in dem Moment, in dem der Trainingsprozess terminiert, wird ein Signal an den Tracker gesendet, der sich vor der Verarbeitung der nächsten Äußerung mit den neu generierten Themenmodellen aktualisiert. Auf diese Weise kann es zwar unter bestimmten Umständen dazu kommen, dass der Klassifikationsprozess mit veralteten Trainingsdaten arbeitet. Dies ist jedoch während der unten beschriebenen Experimente nicht vorgekommen, sondern tritt erst bei großen Trainingsdatenmengen auf. Solche Trainingsdatenmengen sollten jedoch schon selbst genügend Information beinhalten, um gute Ergebnisse bei der Klassifikation zu gewährleisten.

Das SystemInformation-Modul stellt ein generisches *interface* zur Ausgabe von Themeninformationen an das System dar. In zukünftigen Implementierungen kann dieses Modul durch einen ESV-Kanal ersetzt werden. Im Rahmen des Prototypensystems konnte das Modul direkt den Sprachausgabeserver ansteuern, so dass es selbst Antworten auf Themenanfragen durch den Benutzer geben konnte (siehe Abschnitt 6.6). Diese Anfragen wurden vom Dialogmodul/Sprachverstehensmodul erkannt und an das SystemInformation-Modul weitergeleitet.

Die Antworten des SystemInformation-Moduls auf Benutzerfragen nach dem aktuellen Thema unterschieden sich, je nachdem ob kein, ein implizit benanntes oder ein explizit benanntes Thema aktuell erkannt wurde. Der folgende Abschnitt untersucht dies im Detail.

6.5 Implementierung expliziter und impliziter Themennamen

Im Kontext der in Abschnitt 6.6 dargestellten Versuche, aber auch als allgemeine Anforderung stellte sich die Aufgabe, den erkannten Themen Namen zuzuweisen. Ziel dieser Namensgebung war vor allem, für die Benutzer natürlich klingende Antworten auf die Frage nach dem aktuellen Thema (z.B. „*what are we talking about*“) zu finden.

Zur Lösung dieser Aufgabe wurde zwischen zwei Arten von Themennamen – impliziten und expliziten – unterschieden.

6.5.1 Implizite Themennamen

Implizite Themennamen werden vom System selbst gebildet. Prinzipiell stand hierfür die Möglichkeit offen, die Themennamen anhand von (Welt-)Wissen über Situationen zu generieren, so dass z.B. das Vorkommen eines Herdes in einem Themencluster zu der Benennung „Küche“ führt. Dies läuft jedoch der Konzeption des Themenerkennungssystems zuwider, welches ja ausschließlich über Wissen verfügen soll, welches während (vergangener) Kommunikationssituationen mit dem angenommenen Benutzer des Roboters erworben wurde. Aus diesem Grund wurde eine andere Herangehensweise verfolgt: Der Name eines Themas wird über ein besonders prominentes Objekt innerhalb des Themas gebildet. Es bestand die Möglichkeit solche Objekte über ihre Nähe zum Themencluster innerhalb des semantischen Raumes zu bestimmen, auf eine solche Vorgehensweise wurde jedoch verzichtet. Da während der Erstellung des BITT-Korpus auffiel, dass die Versuchspersonen häufig prominente Objekte am Anfang eines Themas erwähnen – so z.B. die Nennung des Schreibtisches, bevor die auf dem Schreibtisch liegenden Objekte benannt werden – wurde folgender Algorithmus implementiert:

Die Objektreferenz[9], welche sich am häufigsten am Anfang von Kommunikationssegmenten befindet, die zu einem Themencluster gehören, liefert den Namen[10] des Clusters/Themas. Die Bestimmung der zu einem Thema gehörenden Kommunikationssegmente führt jedoch zu dem Problem, dass nach jeder Neubestimmung der Themen in Folge einer Aktualisierung des semantischen Raumes alle Kommunikationssegmente erneut den potentiell neu gebildeten Themen zugeordnet werden müssten. Dies wäre nur durch ein komplettes erneutes Tracken aller vergangenen Kommunikationssegmente machbar. Um dies zu vermeiden, wurde ein indirekter Algorithmus verwendet: Für jedes Symbol der Trainingsdatenmenge wird das Symbol bestimmt, welches am häufigsten am Anfang von Kommunikationssegmenten, die das erste Symbol enthalten, stand. Dieser Prozess kann durch einfaches Aufsummieren während der Laufzeit ausgeführt werden.

Werden neue Themencluster bestimmt, wird für alle Symbole des Themenclusters[11] das prominente Symbol bestimmt. Dazu werden für alle im Themencluster vorkom-

[9] Nur Objektreferenzen werden in der Implementierung als Themennamen gehandhabt.

[10] Im Rahmen einer Antwort auf die Frage „*what are we talking about?*“ entspricht der Name dem Plural der Objektbezeichnung. Auf diese Weise werden Antworten wie „*we talk about sofas*“ möglich. Diffizilere Unterscheidungen sind möglich, wurden aber nicht implementiert.

[11] Zur Erinnerung: Obwohl für alle Symbole der Trainingsdatenmenge Distanzen zu allen Clustern bestimmt werden, sind die Symbole eines Themenclusters diejenigen, die der agglomerative Clusteralgorithmus zu dem Cluster zusammengefasst hat.

menden Symbole die Summe der Häufigkeit des ermittelten potentiellen Themennamen mit der durchschnittlichen Korrelation des Symbols zum Thema multipliziert und aufsummiert. Das potentiell themenbenennende Symbol mit der größten Summe dient dann zur Generierung des Themennamens für Anfragen.

Dieser Algorithmus deckt den Fall nicht ab, in dem das Tracking nicht in der Lage war, ein Thema (und damit ein zugehöriges Themencluster) zu bestimmen. Direkt nach einem Themenwechsel ist dies gewünscht; BIRON würde in diesem Fall die Antwort „*sorry, I don't know what we are talking about*“ geben. In Fällen jedoch, in denen ein dem Themenerkennungssystem unbekanntes (z.B. weil neues) Thema initiiert wurde, sollte das Themenerkennungssystem in der Lage sein, eine passende Antwort zu geben. Aus diesem Grund wurde in das System der so genannte **Fokus** mit aufgenommen. Der Fokus enthält Informationen über das aktuelle, nicht abgeschlossene Kommunikationssegment. In Analogie zum skizzierten Algorithmus kann mit Hilfe des Fokus in Fällen, in denen ein unbekanntes Thema vorliegt, die prominente Objektreferenz ermittelt und wie gehabt zur Bildung eines Namens herangezogen werden.

6.5.2 Explizite Themennamen

Ein weiteres Feature des Prototypen besteht in der Möglichkeit für Benutzer, explizite Themennamen anzugeben. Ein Beispiel dafür wäre der Satz „*now we talk about preparing tea*“. Das Sprachverstehensmodul würde an dieser Stelle erkennen, dass der Satz einer expliziten Themensetzung dient und würde das ermittelte Thema „*preparing tea*“ an das Themenerkennungsmodul weitergeben, welches das genannte Thema bis zum nächsten erkannten Themenwechsel als aktives Thema in dem Fokus speichert. Im Gegensatz zu impliziten Themennamen können potentiell beliebige – von der Spracherkennung und dem Sprachverstehen als Themennamen erkannte – Strings als Themennamen dienen.

In der Ermittlung des Themennamens eines Themenclusters werden explizite Themennamen mit Priorität behandelt. Genau wie im Falle impliziter Themennamen werden Symbole als einem Themennamen zugehörig gekennzeichnet, wobei allerdings explizite Themennamen vorrangig – unabhängig von der relativen Häufigkeit – selektiert werden. Ein aktiver, expliziter Themenname substituiert also nicht nur das prominente Symbol eines Kommunikationssegments, sondern verdrängt in der Zählung der Häufigkeit prominenter Symbole alle impliziten Themennamen. Diese aggressive Strategie ist nötig, da nicht davon auszugehen ist, dass Kommunikationspartner regelmäßig den Namen des aktuellen Themas bei Initiierung einer Diskussion dieses Themas benennen. Die Zuordnung von Themennamen zu Themenclustern geschieht wie beschrieben.

Die von dem Prototypen generierten Antworten auf die Frage „*what are we talking about?*“ unterscheiden sich in Hinsicht auf explizite wie implizite Themen. Im Fall impliziter Themen gibt BIRON die vorsichtige Antwort „*it has something to do with...*“, im Fall expliziter Themen äußert er die Antwort „*we talk about...*“.

Die Möglichkeit, implizite wie auch explizite Themenamen anzufragen, stellt eine wichtige Grundlage der im Folgenden beschriebenen experimentellen Evaluation dar, da auf diese Weise das von BIRON erlernte Wissen über Themenzusammenhänge dem Kommunikationspartner direkt zugänglich gemacht werden kann.

6.6 Experimentelle Evaluation

Ein wichtiger Schritt im Kontext der Entwicklung und Implementierung des Themenerkennungssystems bestand darin zu zeigen, dass das System nicht nur *offline* und auf mittelgroßen Trainingsdatenmengen funktioniert, sondern auch *online* und ohne Vortraining im Kontext situierter Kommunikation. Um dies zu erreichen und um weitere Informationen zur Verbesserung des Systems zu gewinnen, wurden mit dem Themenerkennungssystem auf BIRON Experimente durchgeführt.

Im Kontext dieser Experimente wurden insbesondere folgende Aspekte des Systems gemessen:

(i) die quantitative Leistungsfähigkeit des Systems bezüglich des vom Programmierer erwarteten Verhaltens

(ii) die Fähigkeit des Systems, Objekte sinnvoll und *online* Themengruppen zuzuordnen

(iii) die subjektive Zufriedenheit der Benutzer mit BIRONs Fähigkeit der Themenerkennung

Während in (i) gemessen werden sollte, ob das System während der Laufzeit Fehler – z.B. durch zu lange Berechnungszeiten oder Programmierfehler, etc. – produzierte, sollte in (ii) und (iii) beobachtet werden, ob die Modellierung der Themenerkennung als solche sinnvoll ist. Auf diese Weise ließen sich mögliche sinnvolle Erweiterungen des Systems in Hinblick auf eine verbesserte Benutzerfreundlichkeit planen. Im Abschnitt „Ergebnisse“ werde ich detailliert auf die jeweiligen Fragestellungen und Ergebnisse eingehen, die folgenden Unterkapitel widmen sich zunächst dem Aufbau und der Durchführung des Experiments.

Systemkonfiguration

Im Zentrum der Evaluation sollten wie beschrieben der Themenerkennungsalgorithmus sowie die Angemessenheit des Algorithmus für situierte Kommunikationssituationen stehen. Weniger relevant waren

- die Lösung des Segmentierungsproblems. Es war allerdings wichtig, eine natürlich anmutende Segmentierung für das Experiment zu finden. Für das Experiment wurde daher die Standardsegmentierung (s.o.) gewählt - insbesondere die Trennung aufgrund von Phasen des fehlenden Blickkontakts zu BIRON zählten als Themenwechselmarkierung.

- die Fähigkeit des Algorithmus, Fehler anderer Komponenten zu kompensieren.

Da die Objekterkennung eine störende Eingewöhnungsphase der Versuchspersonen notwendig gemacht hätte und zum aktuellen Zeitpunkt auch keine Objektwiedererkennung zur Verfügung stand[12], wurde sie deaktiviert. Aus diesem Grund war es notwendig, die Objekte ausschließlich über ihre Bezeichnungen zu kennzeichnen. Da die Spracherkennung ebenfalls aufgrund der zu dem damaligen Zeitpunkt extrem hohen

[12] vgl. Abschnitt 6.2

Wortfehlerrate simuliert werden musste, wurden beide Probleme auf dieselbe Weise gelöst: Während der Experimente wurden die Äußerungen der Versuchspersonen von einem *Wizard-of-Oz* per Tastatur in das Spracherkennungssystem eingegeben. Objektbezeichnungen, die sich entweder nicht in dem für das Experiment definierten Lexikon befanden oder die nachträglich von den Versuchspersonen geändert wurden, wurden dabei manuell auf eine generische Bezeichnung abgebildet, um sie trotzdem behandeln zu können. Diese Vorgehensweise war legitim, da sie die Effekte einer Objekterkennung simulierte. Zudem traten beide Fälle nur selten ein.

Die Versuchspersonen wurden während des Experiments nicht über die simulierten Prozesse informiert, es handelte sich also wie beschrieben um ein *Wizard-Of-Oz*-Szenario, bei dem die zu evaluierenden Komponenten natürlich nicht simuliert wurden. Aus wissenschaftsethischen Gründen wurden die Versuchspersonen nach dem Experiment über den Anteil an simulierten Roboterfähigkeiten informiert.

Für das Experiment war weiterhin relevant, dass BIRON über kein Vorwissen über Themengruppen verfügte - jegliches thematische Wissen sollte ausschließlich während des Versuchs selbst gewonnen werden. Auf diese Weise wurde im Gegensatz zu der *offline*-Evaluation ein Fokus auf die Echtzeitfähigkeiten des Systems gelegt. Vor jedem Experiment und Experimentabschnitt wurde der Speicher des Themenerkennungssystems gelöscht.

Für das Experiment wurde ein semantischer Raum nach Rieger verwendet. Das Distanzmaß war der Korrelationskoeffizient, das Limit für den agglomerativen Clusteralgorithmus wurde auf 0.01 gesetzt. Die Segmentierung erfolgte somit dynamisch, es wurde im Gegensatz zu den *offline*-Experimenten keine Clusteranzahl bzw. kein Clustermaximum spezifiziert.

Experimentaufbau

Wie aus Abbildung 6.8 ersichtlich[13], kommunizierten die Versuchspersonen mit BIRON über einzelne Objekte, die sie von einem hinter ihnen befindlichen Tisch nehmen konnten. Auf dem Tisch vor den Versuchspersonen waren drei Flächen markiert, auf die Objekte, über die kommuniziert werden sollte, abgelegt werden konnten. Auf diese Weise konnten die Versuchspersonen nie über mehr als drei Objekte gleichzeitig kommunizieren.

Die Objekte wurden so ausgewählt, dass sich vier Themengruppen von je fünf Objekten erkennen ließen. Zwischen den thematischen Zuordnungen existierten jedoch mögliche Unschärfen, so konnte eine Keksdose sowohl thematisch mit anderen Süßigkeiten in Verbindung gebracht werden, aber auch mit einem Teegeschirr. Einen Überblick über die im Versuchsaufbau befindlichen Objekte bietet Grafik 6.9. Im Anhang (vgl. Abschnitt 8.4 auf Seite 181) findet sich eine genaue Auflistung der Objekte, der erwarteten Objektnamen und der erwarteten Themengruppenbezeichnungen.

Die Experimente wurden mittels eines externen Camcorders und eines Raummikrofons aufgezeichnet.

Ablauf

Das Experiment gliederte sich in zwei Phasen:

[13] Das Bild ist gestellt; der Versuchsaufbau ist allerdings mit dem in den Experimenten verwendeten identisch. Üblicherweise wurde mehr als ein Objekt gleichzeitig gezeigt.

Abbildung 6.8: Experimentaufbau der Online-Experimente

1. Freie Kommunikation
2. Hauptteil

Der erste Experimentteil diente der Gewöhnung der Versuchspersonen an die Kommunikation mit BIRON sowie der Gewinnung von frei geäußerten Sätzen, die der separaten Evaluierung des Sprachverstehens (ASU) dienen sollten. Zur Anleitung wurden die Personen durch einen Anweisungsbogen (Abschnitt 8.5 auf Seite 182f) dazu aufgefordert, BIRON Objekte zu zeigen, wobei sie nicht explizit auf die Möglichkeit, Themen zu bezeichnen oder zu erfragen, hingewiesen wurden. Die Versuchspersonen füllten im Anschluss einen Fragebogen aus (Abschnitt 8.6 auf Seite 185f).

In einem folgenden Abschnitt wurden die Versuchspersonen über ihre eigentliche Aufgabe informiert. Dies geschah in zwei Schritten: Erstens rezipierten die Versuchspersonen den zweiten Einführungsbogen (Abschnitt 8.5 auf Seite 182ff), zweitens wurde das korrekte Verhalten (z.B. die richtige Positionierung für eine optimale Erkennung des Beinpaares durch den Laserscanner) während des zweiten Experimentteils durch den Versuchsleiter vorgeführt. Dies stellte sich nach Vorversuchen als nützlich heraus, da viele Versuchspersonen im Vorfeld die Anweisungen auf dem Anleitungsbogen nur unvollständig oder teilweise falsch befolgten.

Im Rahmen des zweiten Experimentteils hatten die Versuchspersonen die Aufgabe, BIRON thematisch zusammenhängende Objekte zu zeigen. Sie sollten zusammengehörende Objekte auf den Tisch legen, sie BIRON zeigen und nach Wahl das Thema benennen bzw. nach dem Thema fragen. Wie aus dem Fragebogen ersichtlich

Abbildung 6.9: Im *online*-Experiment verwendete Objekte

wurden den Versuchspersonen dazu feste Äußerungen an die Hand gegeben, um Fehler im Bereich des Sprachverstehens auszuschießen. Die Äußerungen waren dabei im Speziellen:

- „*hello BIRON*“ – Begrüßung
- „*this is a(n)* <OBJEKTNAME>“ – Zeigen von Objekten
- „*these are* <OBJEKTNAME>“ – Zeigen von Objekten
- „*we talk about* <THEMENNAME>“ – Benennen eines Themennamens
- „*what are we talking about*“ – Abfragen des aktuellen Themas

Den Versuchspersonen wurde keine Reihenfolge der Äußerungen vorgegeben, abgesehen von dem Beginn durch die Begrüßung. Abbildung 6.10 auf der nächsten Seite gibt einen Eindruck eines Experimentablaufs.

Der zweite Experimentteil wurde durch die Versuchspersonen selbst terminiert: Nach sechs Minuten wies der Versuchsleiter die Versuchspersonen darauf hin, nach eigenem Ermessen das Experiment zu beenden. Auf diese Weise konnte eine Minimallänge des Experiments festgelegt werden, es wurde den Versuchspersonen aber ermöglicht, das System nach ihren eigenen Vorstellungen ohne Einschränkungen durch harte Zeitgrenzen zu testen. Tabelle 6.1 zeigt die resultierende Dauer der Experimente.

VP: hello biron
BIRON: hi, hello
VP: (legt Elefant, Kuh und Ente auf den Tisch vor sich)
VP: (zeigt) this is a duck
BIRON: it's very nice
VP: (zeigt) this is an elephant
BIRON: oh I really like it
VP: (zeigt) this is a cow
BIRON: it's really interesting
VP: what are we talking about
BIRON: it has something to do with the duck
VP: these are animals
BIRON: OK
VP: what are we talking about
BIRON: we talk about animals
VP: (Tauscht die Objekte gegen andere aus)
(...)

Abbildung 6.10: Beispiel für einen Dialog mit BIRON im Rahmen des *online*-Experiments

	Gesamtdauer	**Durchschnittliche Dauer**
Hauptgruppe	91m 17s	8m 15s
Kontrollgruppe	45m 4s	7m 5s
Gesamt	136m 21s	8m 0s

Tabelle 6.1: Experimentdauer

Im Anschluss an den zweiten Experimentteil füllten die Versuchspersonen wiederum einen Fragebogen aus (siehe Abschnitt 8.6).

Ergebnisse

An dem Experiment nahmen 19 (12m/7w) Versuchspersonen teil. Zwei Experimente wurden aufgrund von technischen Problemen mit dem Robotersystem oder mit den Aufzeichnungsgeräten als ungültig erklärt und konnten aufgrund zeitlicher und inhaltlicher Beschränkungen nicht wiederholt werden, so dass an dem Experiment real 17 Versuchspersonen teilnahmen. Sechs Versuchspersonen waren dabei Teil einer Kontrollgruppe, die im zweiten Experimentteil nicht die Möglichkeit hatten, das aktuelle Thema zu erfragen. Auf diese Weise sollte gemessen werden, wie sehr die Versuchspersonen BIRON die Fähigkeit, Themen zu erkennen, zubilligten, ohne diese Fähigkeit getestet zu haben.

Wie anfangs[14] dargestellt, sollten verschiedene Aspekte der Themenerkennung gemessen werden. Die folgenden Abschnitte stellen die Art der Messung sowie die Ergebnisse dar.

[14] vgl. S.148

Nr.	Korrekt Implizit	Korrekt Explizit	Korrekt Gesamt	Falsch Gesamt	Anteil korrekt
1	2	6	8	0	1
2	4	6	10	0	1
3	7	2	9	3	0.75
4	3	5	8	3	0.73
5	6	4	10	1	0.91
6	2	6	8	2	0.8
7	5	7	12	0	1
8	2	4	6	3	0.66
9	6	7	13	0	1.0
10	2	7	9	1	0.9
11	4	7	11	0	1
Gesamt:	43	61	104	13	0.89

Tabelle 6.2: Anzahl korrekter/inkorrekter Antworten von BIRON

Quantitative Leistungsfähigkeit Um die quantitative Leistungsfähigkeit des Systems zu ermitteln, wurde jede der Antworten BIRONs auf Benutzerfragen nach dem aktuellen Thema („*what are we talking about?*“) betrachtet. Tabelle 6.2 gibt Auskunft über die Ergebnisse. Für diese Auswertung wurden selbstverständlich nur die elf Experimente der Hauptgruppe herangezogen, da die Versuchspersonen in der Kontrollgruppe keine Möglichkeit hatten, das aktuelle Thema zu erfragen.

Die korrekten Antworten von BIRON wurden in implizit korrekt und explizit korrekt unterteilt. Implizit korrekt bedeutet, dass BIRON zwar keinen Themennamen vom Benutzer genannt bekommen hatte, aber trotzdem korrekt antwortete. Eine korrekte Antwort in diesem Fall bestand in der Angabe des prominenten Objektes für das Thema, also des erstgenannten. Bei den falschen Antworten von BIRON war eine solche Unterteilung nicht sinnvoll, da es mehr oder minder zufällig war, ob die falsche Antwort implizit, explizit oder sogar „weder noch“ – im Falle eines „*I don't know what we are talking about*“ – war.

In Tabelle 6.3 sind die Fehlergründe aufgeführt. Aufgrund der niedrigen Zahlen sind vergleichende statistisch signifikante Aussagen zwischen den Fehlerquellen zwar unmöglich, aber die Gesamtzahl an Fehlern in Relation zu den korrekten Fällen zeigt eindeutig, dass das Themenerkennungssystem als solches zumindest weitgehend fehlerfrei funktioniert.

Fehlerquelle	Anzahl
Segmentierungsfehler	8
Mehrfachthemen	3
Unbekannt	2
Gesamt	13

Tabelle 6.3: Fehler nach Fehlerquellen

Die Bewertung, dass eine Roboterantwort auf eine Frage korrekt oder falsch ist, wurde von dem Versuchsleiter anhand der aufgezeichneten Videodaten im Nachhinein

bestimmt. Dabei floss aber das Verhalten der Versuchspersonen stark mit ein, wie die Existenz von Mehrfachthemen-Fehlern zeigt (siehe unten). Im Falle einer mimisch oder verbal geäußerten Zurückweisung einer Roboterantwort durch eine Versuchsperson wurde folglich in jedem Fall ein Fehler notiert. Jede Roboterantwort wurde unweigerlich als korrekt oder falsch klassifiziert.

Die **Segmentierungsfehler** stellen den größten Anteil der Fehler. Zwar hat Wegsehen als Themengrenzenindikator im Rahmen dieses Experiments gut funktioniert, ist aber für zukünftige Anwendungen wahrscheinlich zu strikt, da Personen und Roboter während einer Kommunikation nicht unentwegt Blickkontakt halten. Fehler traten sowohl durch Unter- als auch durch Übersegmentierungen auf.

Das Kürzel **Mehrfachthemen** steht für Fehler, die aufgetreten sind, weil der Benutzer ein Objekt gerne zwei verschiedenen Themen zugeordnet gesehen hätte (z.B. die Kekse, die sowohl dem Süßigkeitenthema als auch dem Teekoch-Thema zugeordnet wurden).

Unbekannte Fehlerursachen traten auf, ohne dass die Fehlerquelle rekonstruiert werden konnte.

Subjektive Leistung Die subjektive Leistung von BIRON, Themenzusammenhänge zu erkennen, wurde in zwei verschiedenen Testverfahren gemessen. Zentral für diese Messungen war die Frage

Glauben Sie dass BIRON in der Lage war, im Dialog Zusammenhänge zu erkennen? (Ja / Größtenteils schon / Größtenteils nicht/ Nein)

die nach der Eingewöhnungsphase gestellt wurde (Fragebogen 1). Sowohl die Kontrollgruppe als auch die Hauptgruppe wurde nach dem zweiten Experimentteil mit folgender Aufgabe konfrontiert:

Als wie gut beurteilen Sie die Fähigkeit von BIRON, Zusammenhänge zu erkennen? (Sehr gut / Gut / Eher schlecht / Schlecht)

Dabei ist zu bedenken, dass die zweite Frage schärfer gestellt ist als die erste – man kann BIRON eine Fähigkeit, Zusammenhänge zu erkennen zugestehen, auch wenn diese schlechte Ergebnisse erzeugt.

Anhand einer weiteren Frage, die aber nicht statistisch-kontrastiv behandelt wurde, konnte gemessen werden, wie zufrieden die Versuchspersonen mit BIRONs allgemeiner Leistung waren:

Fanden Sie, dass BIRON seine Aufgaben gut bewältigt hat? (Ja, immer / Meistens / Manchmal / Nein, nie)

Auch hier lässt sich ein überwiegend positiver Eindruck erkennen; vier Personen haben mit „sehr gut“, elf mit „gut“ und nur zwei mit „eher schlecht“ geantwortet.

Tabelle 6.4 stellt die Ergebnisse im Zusammenhang dar. Dabei sind der Einfachheit halber die Antworten durch Symbole dargestellt: sehr gut/ja, immer: ++; gut/meistens: +; eher schlecht/manchmal: -; schlecht/nein, nie: --.

In Bezug auf die Fragen nach BIRONs Fähigkeit, Zusammenhänge zu erkennen, wurden zwei Aspekte statistisch untersucht:

1. Änderungen in den Antworten vor und nach dem zweiten Experimentteil in der Hauptgruppe, und

Nr.	Zusammenhänge nach 1. Versuchsteil	Zusammenhänge nach 2. Versuchsteil	Aufgabe	Gruppen
1	-	+	++	++
2	+	+	++	++
3	-	+	+	+
4	-	+	-	++
5	+	+	+	+
6	-	+	+	+
7	+	+	+	++
8	-	+	+	++
9	-	+	++	+
10	+	+	++	++
11	-	-	+	++
K 1	+	-	+	++
K 2	-	-	+	+
K 3	-	-	+	++
K 4	- -	-	-	++
K 5	-	-	+	++
K 6	-	-	+	++

Tabelle 6.4: Benutzerbeurteilungen

2. Unterschiede in den Antworten der Haupt- und der Kontrollgruppe nach dem zweiten Experimentteil.

Beide Kontrastierungen dienen dem Zweck zu erkennen, in wie weit die Versuchspersonen BIRON die Fähigkeit, Zusammenhänge zu erkennen, zuschreiben, ohne dafür eine Grundlage zu haben. Unterscheiden sich diese Einschätzungen signifikant (zum positiven) von den Antworten der Hauptgruppe, nachdem sie die Themenerkennung haben testen können, kann dem System zumindest die Simulation der Fähigkeit, Zusammenhänge zu erkennen, nicht abgesprochen werden.

Die Vorher-nachher-Analyse innerhalb der Hauptgruppe wurde mit dem Vorzeichen-Rangsummentest nach Wilcoxon durchgeführt (Bortz und Lienert, 2003). Das zweiseitige Testverfahren zeigte einen signifikanten Unterschied (Signifikanzniveau von 5%) zwischen den Antworten vor und den Antworten nach dem zweiten Experimentteil. Der U-Rangsummentest nach Mann, Whitney und Wilcoxon (ebd.) belegte einen signifikanten, zweiseitigen Unterschied (2%) zwischen der Hauptgruppe und der Kontrollgruppe nach dem zweiten Experimentteil.

Themengruppenbildung Um die Qualität der gebildeten Themengruppen zu messen, wurden die Versuchspersonen am Ende des Experiments – also nach dem Ausfüllen des zweiten Fragebogens – gebeten, die Themengruppen als *sehr gut/gut/eher schlecht/schlecht* zu bewerten[15]. Dazu wurden den Versuchspersonen die vom Themenerkennungssystem gebildeten Wortcluster gezeigt (z.B. „*mouse keyboard*

[15]Die exakte Fragestellung lautete: „Lassen Sie sich die von BIRON gebildeten Themengruppen zeigen. Als wie gut bewerten Sie sie?“

memorystick“). Für diese Auswertung wurden die Versuchspersonen der Hauptgruppe so wie der Kontrollgruppe hinzugezogen.

In zwölf Fällen wurden die von dem Themenerkennungssystem gebildeten Gruppen als „sehr gut“ bezeichnet, in fünf Fällen als „gut“ (vgl. Tabelle 6.4). Keine der Versuchpersonen gab „eher schlecht“ oder „schlecht“ als Antwort an. Es ließ sich wie erwartet kein signifikanter statistischer Unterschied zwischen der Haupt- und der Kontrollgruppe ermitteln.

Diskussion

Im Rahmen der Evaluation des *online*-Experiments wurden wie beschrieben verschiedene Analysen vorgenommen: Eine Analyse des Anteils korrekter Antworten auf Benutzerfragen nach dem aktuellen Thema, eine statistisch-kontrastive Analyse der Benutzerzufriedenheit mit und ohne Themenerkennung so wie eine nicht-kontrastive Untersuchung der Benutzerzufriedenheit mit bestimmten Aspekten des Systems.

Die quantitative Analyse des Anteils korrekter Antworten von BIRON zeigt eindeutig, dass das System konzeptuell wie real in einer sehr großen Anzahl von Fällen wie gewünscht funktioniert. Die quantitative Analyse der korrekten bzw. inkorrekten Antworten von BIRON zeigt noch einen gewissen Verbesserungsspielraum insbesondere bezüglich der Segmentierung. Angesichts des optimierten Szenarios für die dargestellten Experimente kann vorerst davon ausgegangen werden, dass es sich hierbei um ein „hartes“ Problem handelt, auf das sich zukünftige Forschungen konzentrieren sollten, wenn sie den vorgestellten Ansatz weiterverfolgen. Eine weitere Optimierung des Segmentierungsprozesses war aber im Rahmen des Experiments nicht sinnvoll. Auch für weitergehende Ansätze prinzipiell problematisch ist natürlich, dass bei einem frühen Trainingsstand nahezu jegliche Segmentierungsfehler sofort in Themenerkennungsfehler umgewandelt werden, erst später entsteht durch eine breitere Trainingsbasis eine gewisse Fehlerresistenz.

Das Experiment hat eindeutig gezeigt, dass die Versuchspersonen BIRON mit dem Themenerkennungssystem die Fähigkeit, Zusammenhänge zu erkennen, zusprechen und mit den gebildeten Themengruppen zufrieden sind. Die Analyse der subjektiven Einschätzung kann somit als voller Erfolg gewertet werden. Kontrastiert man die positivere Bewertung der gebildeten Themengruppen mit der Bewertung von BIRONs Fähigkeit, seine Aufgabe zu erfüllen, so ist erkennbar, dass weniger das Kernverfahren, sondern vielmehr andere Aspekte der Einbindung der Themenerkennung an das Gesamtsystem verbessert werden müssen. Im folgenden Abschnitt möchte ich eine solche (mögliche) Verbesserung diskutieren.

Erwartete joint attention Wie beschrieben ist der klare Unterschied zwischen den Antworten auf die Frage nach BIRONS Fähigkeit, Zusammenhänge zu erkennen – die fast immer mit „gut“, aber nie mit „sehr gut“ beantwortet wurde – und der Frage nach der Einschätzung der Qualität der Themencluster, die in mehr als $\frac{2}{3}$ aller Fälle mit „sehr gut“ beantwortet wurde, auffällig. Auch wenn sich die Fragen inhaltlich unterscheiden, liegt hier eine interessante Diskrepanz vor. Meiner Ansicht nach ist dies Verhalten durch Erwartungen zu erklären, die BIRON während der Kommunikation nicht erfüllt hat. Die Versuchspersonen fragten im ersten Experimentteil oft „*what is this?*“ und zeigten dabei auf ein Objekt. Sie erwarteten also von BIRON eine bessere

joint attention und die Fähigkeit, Objekte von sich aus zu benennen und zu erkennen. Dies setzte sich im zweiten Experimentteil dahingehend fort, als dass die Versuchspersonen teilweise erwarteten, dass BIRON die vor ihm liegenden Objekte und das dazugehörige Thema ohne Nennung der Objekte durch den Benutzer erkennen sollte. Als Fazit sollte der Roboter also nicht nur in der Lage sein, benannte Objekte zu erkennen, sondern alle Objekte, die einer *joint attention*-Situation gehören, und diese zur Themenerkennung und zum Training derselben hinzuziehen.

Diese Erkenntnis ist von großer Wichtigkeit für eine situierte Themenerkennung auf einem mobilen Robotersystem, da sie auch im Zusammenhang mit einem grundsätzlichen Problem von dialogischer Kommunikation steht (Nakata u. a., 2002): Die Erwähnung von relevanten Objekten geschieht oftmals nur initial und folgend anaphorisch oder gar nicht, weswegen die in einem semantischen Raum resultierenden Satzvektoren geringe Häufigkeiten relevanter Symbole enthalten. Selbst wenn anaphorische Referenzen aufgelöst werden können, würde die Erkennung von Objekten, die für die momentane Handlung wichtig sind, mit großer Wahrscheinlichkeit zu wesentlich reicheren Trainingsdaten führen.

Strategie der impliziten Antworten Eine weitere Erkenntnis ist, dass die Versuchspersonen mit den von BIRON gegebenen, impliziten Antworten oftmals unzufrieden waren. Dies war daran erkennbar, dass in vielen Fällen auf eine solche Antwort eine Konkretisierung durch die Versuchspersonen folgte (z.B. „*we talk about animals*“). Die gewählte Strategie, das erste Objekt als Themenbezeichner heranzuziehen, konnte im Kontext dieses Experiments als nicht erfolgreich angesehen werden. Dies beruht vermutlich im Wesentlichen darauf, dass innerhalb der Objektgruppen keine prominenten Objekte, die als Themennamen fungieren konnten, vorhanden waren. In wiefern sich die Auswahl eines prominenten Objektes generell zur Namensgebung eines nicht explizit benannten Themas eignet – oder ob gegebenenfalls zwangsläufig doch auf Weltwissen zurückgegriffen werden muss – bleibt abzuwarten.

Insgesamt möchte ich an dieser Stelle festhalten, dass das Experiment trotz der vielen vereinfachenden Bedingungen als Erfolg gewertet werden kann. Der Themenerkennungsalgorithmus funktioniert in situierter Kommunikation in Echtzeit. Weiterhin hat der vorgestellte Algorithmus des großen Vorteil, dass er im Kontrast zu möglichen, einfacheren Algorithmen, mit denen sich ggf. ein vergleichbares Ergebnis hätte erzielen lassen können, auch auf größeren Trainingsdatenmengen skaliert, wie in den *offline*-Untersuchungen gezeigt werden konnte.

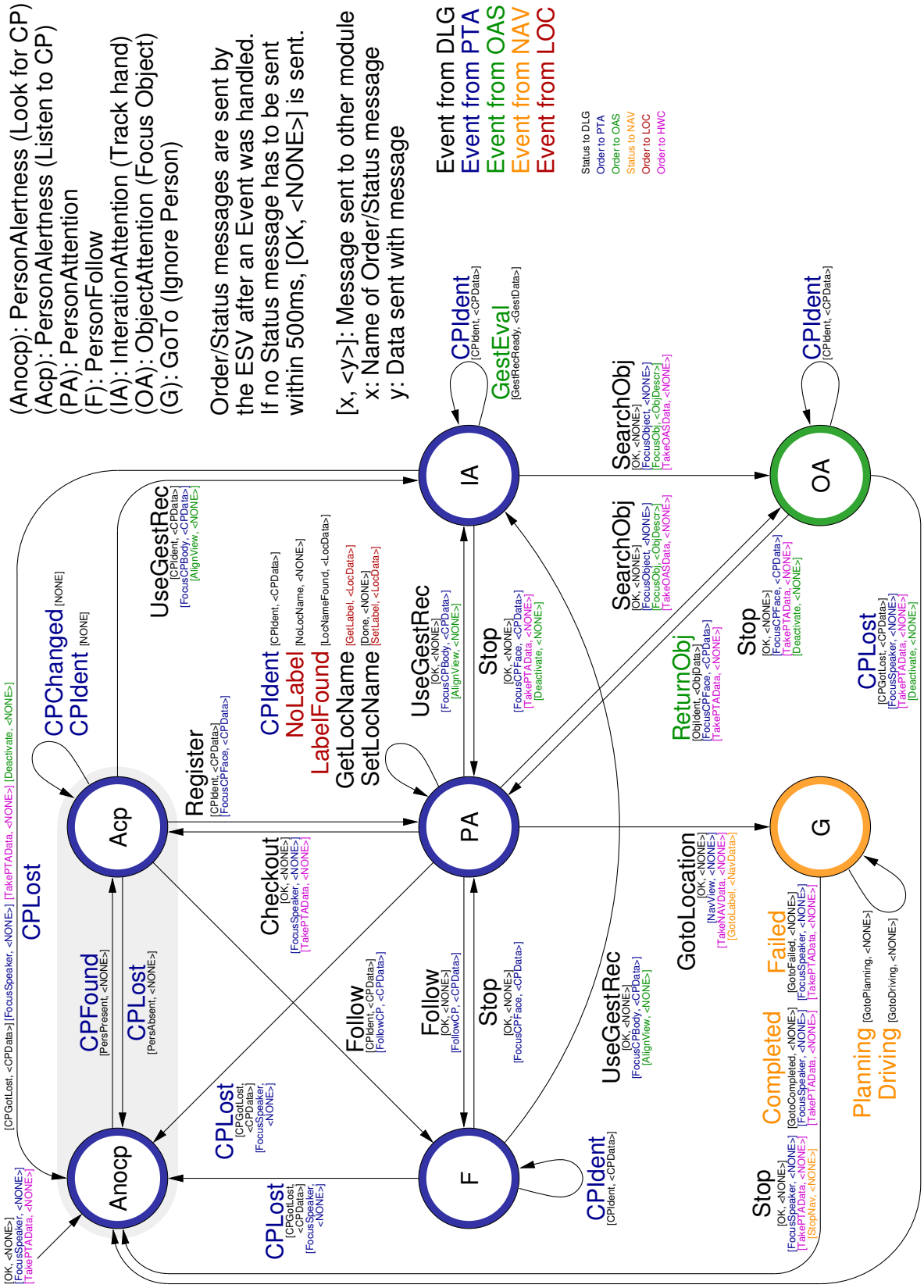

Abbildung 6.11: Zustände und Transitionen des ESV

7 Diskussion und Ausblick

Das Ziel dieser Arbeit war es, die Möglichkeiten von sozial-natürlichsprachlich kommunizierenden (Haushalts-)Robotern durch die Entwicklung eines multimodal operierenden, dynamischen, situiert arbeitenden und *online*-fähigen Themenerkennungssystems zu erweitern. Dabei wurde insbesondere auf Vorarbeiten aus der statistischen Textanalyse zurückgegriffen, da in diesem Gebiet schon eine große Anzahl effektiver und effizienter Ansätze zur Themenerkennung existiert.

Ein Themenerkennungssystem kann für ein natürlichsprachlich interagierendes Robotersystem eine Vielzahl von Möglichkeiten eröffnen. Beispiele hierfür sind das bessere Verständnis situierter Sprache und Verbesserungen in der Wort-Fehler-Rate der Spracherkennung durch die Verwendung thematisch reduzierter Lexika.

Die Umsetzung geschah im Besonderen anhand der Entwicklung eines auf einem Roboter lauffähigen Prototypensystems zur Themenerkennung, das in verschiedenen Experimenten – darunter einer Benutzerstudie mit untrainierten[1] Versuchspersonen – verbessert und mit Erfolg getestet wurde. Die grundsätzlichen – und erfüllten – Anforderungen an den Prototypen bestanden darin, dass die Themenerkennung *online*, also in quasi-Echtzeit ablaufen sollte und dass kein vordefiniertes Weltwissen z.B. in Form vordefinierter Themen oder Themenanzahlen in den Prozess mit einfließen sollte, da nur auf diese Weise dem dynamischen Charakter von Haushaltsumgebungen Genüge getan werden konnte. Die Anforderung einer multimodalen Vorgehensweise ergab sich aus der Komplexität der Aufgabe, die eine rein unimodale Vorgehensweise schwierig bis unmöglich gemacht hätte.

Speziell hervorzuheben ist an dieser Stelle die Entwicklung des für wissenschaftliche Zwecke frei zur Verfügung gestellten BITT-Korpus, in dem 29 Versuchspersonen dem Roboter BIRON in einem an ein *home tour*-Szenario angelehnten Experiment einen thematisch geordneten Raum vorstellen. Dieser Korpus diente als eine wichtige Grundlage für die *offline*-Evaluation des Systems.

Ein besonderer Erfolg des Themenerkennungsalgorithmus besteht in dem Umstand, dass er entgegen den üblichen Anwendungsfällen von statistisch-analytischen Verfahren schon im Kontext extrem kleiner Trainingsdatenmengen – wie sie in den Benutzerstudien entstanden sind – sehr gute Ergebnisse zu liefern in der Lage ist. Dass der Algorithmus seine Fähigkeit, mit größeren Datenmengen umzugehen, dabei nicht eingebüßt hat, konnte in den jeweiligen *offline*-Studien bewiesen werden.

7.1 Zusammenfassung und Diskussion

Ich möchte an dieser Stelle im Überblick die wichtigsten Erkenntnisse dieser Arbeit zusammenfassen.

[1] Im Rahmen des Versuchs fand durchaus eine Einführung in die Kommunikation mit dem Robotersystem statt. „Untrainiert“ bedeutet in diesem Kontext, dass die Personen bisher keine oder nur geringfügige Erfahrungen im Umgang mit BIRON hatten.

Die Evaluation des *offline*-Systems anhand des Korpus zeigte, dass in dem BITT-Korpus – und somit mit großer Wahrscheinlichkeit in *home tour*-Szenarien im Allgemeinen – eine starke Bindung von Themen an Objekte bzw. Objektreferenzen herrscht. Aus diesem Grund konnte mit Hilfe einer multimodalen Erweiterung gängiger semantischer Räume – des einfachen Vektorraummodells, des Fuzzy-Semantics-Raums und der LSA – eine gravierende Verbesserung der Leistung des Systems im Vergleich zu üblichen, rein wortbasierten Ansätzen gewonnen werden. Außerdem konnte ein stark störender Einfluss selten vorkommender Worte/Symbole entdeckt werden, der in anderen Korpora aufgrund der größeren Länge der zu klassifizierenden Entitäten nicht so stark zum Tragen kommt. Zur Kompensation dieses Fehlers wurde der Wert C bzw C_{rel} – also die minimale Auftretenshäufigkeit eines Wortes/Symbols, ggf. relativ zum Korpus – eingeführt, aber auch eine klassische Vorgehensweise durch Anwendung einer LSA konnte das Problem beheben. Auf diese Weise kann auch auf der wünschenswerten Basis von geringen Werten für C – bei denen nur wenige Symbole aus der Datenmenge herausgenommen werden – gearbeitet werden.

Eine wichtige Erkenntnis bestand auch darin, dass zumindest unter optimalen Bedingungen eine automatische Terminierung des Clustervorgangs anhand einer Analyse des Verlaufs der Clusterdistanzen möglich ist. Auf diese Weise kann auch auf mittelgroßen Traningsdatenmengen ohne die Verwendung mehr oder minder willkürlicher Schwellwerte die optimale Themenanzahl gefunden werden. Alternativ kann natürlich trotzdem versucht werden z.B. auf der Basis des Korrelationskoeffizienten als Distanzmaß einen Schwellwert nahe des Wertes 0 (=unkorreliert) zu finden.

Wie beschrieben wurde anschließend eine Implementierung des Systems auf dem Roboter BIRON durchgeführt, bei dem die *online*-Fähigkeit des Ansatzes getestet wurde. Das System wurde dazu um die Fähigkeit der eigenständigen Benennung von Themen, bzw. die der Einbeziehung von Themennamen, die von Benutzern vorgegeben worden waren, erweitert. Diese Erweiterungen stellen einen notwendigen Schritt zur Verwendung des Themenerkennungssystems im Rahmen des Dialogsystems dar.

Im Rahmen von einer Benutzerstudie konnte gezeigt werden, dass das System trotz des global – also stets auf der gesamten Trainingsdatenmenge – arbeitenden Themenerkennungsalgorithmus in Echtzeit das Thema neuer Benutzeräußerungen erkennen kann. Die Verwendung eines global arbeitenden Algorithmus birgt den Vorteil, dass auf diese Weise eine verbesserte Fehlerkorrektur bei Erhalt neuer Trainingsdaten möglich ist.

Eine quantitative Analyse der Reaktionen des Systems auf Themenanfragen ergab, dass im Kontext des Experiments zu ca. 90% korrekte Antworten erzielt wurden. Die häufigste Fehlerursache war dabei eine fehlerhafte Segmentierung nach Themenabschnitten, die zum Training des semantischen Raums dient. Somit ist der Kernalgorithmus – also die Themenerkennung als solche ohne die separate Segmentierung – funktionsfähig.

Die subjektive Einschätzung der Leistungen des Themenerkenners von Seite der Versuchspersonen aus lieferte ebenfalls ein sehr positives Bild: Die Versuchspersonen waren sowohl mit den Antworten des Systems als auch mit der internen Repräsentation der Themen zufrieden. Letztere wurde besser bewertet, was auf einen Verbesserungsspielraum bezüglich der Einbindung der Resultate des Themenerkennungssystems in das Dialogsystem hindeutet.

Zusätzlich wurden die Versuchspersonen gebeten, die Fähigkeit des Systems, (the-

matische) Zusammenhänge zu erkennen, zu bewerten. In einem Vorher-Nachher-Vergleich so wie in Vergleich mit einer Kontrollgruppe ergaben sich statistisch signifikante Differenzen, die ebenfalls die Leistungsfähigkeit des Systems auch und gerade in Bezug auf die Benutzerzufriedenheit belegten. Dieses Ergebnis erlaubt weiterhin den Schluss, dass die Implementierung der Unterstützung vom Benutzer vorgegebener Themennamen erfolgreich war.

7.2 Ausblick

Aus den *online*-Studien war wie erwartet erkennbar, dass gerade im Kontext sehr geringer bzw. nicht vorhandener Trainingsgrundlagen die Genauigkeit der Segmentierung direkten Einfluss auf die Qualität der Themenerkennung hat. Aus diesem Grund sollte bei zukünftigen Adaptionen des entwickelten Algorithmus eine Optimierung der Segmentierung – wie sie im Rahmen dieser Arbeit nicht mehr erfolgen konnte – im Vordergrund stehen. Insbesondere die direkte Einbeziehung von semantischer Information bzw. Diskursinformation, wie sie in sehr grober Weise schon in dieser Arbeit durch die Verarbeitung von Benutzerphrasen der Art „*let's change the topic*“ bzw. „*we talk about...*“ vorgenommen wurde, kann in Zukunft eine sehr hilfreiche Informationsquelle zur Detektion von Themengrenzen sein.

Zukünftige Arbeiten sollten insbesondere auf die Kombination eines erweiterten Szenarios unter Bezugnahme auf eine stärkere Nutzbarmachung von thematischer Information zielen. Das *home tour*-Szenario stellte zwar aus den diskutierten Gründen eine ausgezeichnete Grundlage für die Entwicklung eines Themenerkennungssystems dar, zukünftige Forschung sollte jedoch versuchen, noch näher am „Alltag“ eines Haushaltsroboters zu forschen.

Der Aspekt der stärkeren Nutzbarmachung der thematischen Information steht natürlich mit der stärkeren Anbindung an ein Szenario bzw. an den Alltag eines Haushaltsroboters in direktem Zusammenhang. Grundsätzlich ist die Frage zu stellen, ob die in dieser Arbeit angenommene, nicht-hierarchische Themenstruktur den Anforderungen entspricht. Im Zweifelsfall lässt aber das hier vorgestellte hierarchisch-agglomerative Clusterverfahren die Erkennung hierarchischer Themen zu, so dass zumindest theoretisch einer Erweiterung des Systems in diese Richtung keine Hindernisse im Weg stehen.

In den Rahmen der besseren Nutzbarmachung thematischer Information fällt auch und insbesondere die stärkere Anpassung an Erwartungen und Wünsche von Benutzern. Im Rahmen dieser Arbeit wurde durch die Erweiterung des Systems, die das Lernen von benutzerspezifizierten Themennamen ermöglicht, schon ein Schritt in diese Richtung unternommen. Zusätzlich zu der Assoziation von (expliziten) Themennamen zu Themenclustern können in zukünftigen Forschungen Emotionen oder andere soziale Randbedingungen eingebunden werden, die den Benutzern das Gefühl geben, mit einem empathischen Subjekt und nicht mit einer Maschine zu kommunizieren.

Zur Verbesserung der Effektivität des Themenerkennungssystems wurden im Rahmen dieser Arbeit – neben der Verbesserung des Segmentierung – zwei Wege vorgezeichnet. Zum einen sollte untersucht werden, welche weiteren Symbole verschiedener Modalitäten noch in die Trainingsmenge des semantischen Raumes mit aufgenommen werden können. Ein Beispiel dafür wären Ortskennzeichnungen, zumal BIRON über die Fähigkeit der Lokalisation verfügt (Spexard u. a., 2006). Ein anderes Beispiel

wären emotionale Indikatoren (Hegel u. a., 2006). Weitere Modalitäten sind zweifellos denkbar und möglich.

Zum anderen kann ggf. durch eine Optimierung des Ansatzes der gemischten Modelle – wie im Abschnitt 5.4.1 dargestellt – eine Kombination aus den sehr gut funktionierenden, rein auf Objektreferenzen basierenden semantischen Räumen mit den weniger gut funktionierenden, dafür aber mehr Informationen liefernden gemischten semantischen Räumen erzeugt werden.

Wie beschrieben stellt diese Arbeit den Versuch einer Fusion zweier Forschungsgebiete – der statistischen Themenerkennung und der *interaction robotics* – dar. Diese Arbeit ist somit als interdisziplinär zu betrachten und war somit den Vor- und Nachteilen interdisziplinärer Arbeit ausgesetzt. Eine Eingrenzung der Fragestellung war zwingend erforderlich, da keine durch eine wissenschaftliche Gemeinschaft gegebene strikte Thematik vorgegeben war. Ähnlich dem in dieser Arbeit entwickelten Themenerkennungsprozess musste auch diese Arbeit ihr Thema dynamisch bilden.

Ein Ergebnis dieser Vorgehensweise war die Festlegung auf rein selbstlernende Verfahren, die über keinerlei vordefiniertes Wissen verfügen sollten. Obwohl diese Entscheidung fundiert begründet wurde – die Implementierung von vordefiniertem Wissen scheitert an der hohen Dynamik häuslicher Umgebungen – liegt die Zukunft der Themenerkennung im Bereich der Mensch-Roboter-Kommunikation vermutlich dennoch in einem Mittelweg aus vordefiniertem Wissen – z.B. durch Ontologien – und dynamischen Erkenntnissen. Eine mögliche Implementierung eines solchen Ansatzes wäre die Verwendung von ontologisch-thematischer Information als *default*-Fall, die dann durch dynamisch generiertes Wissen überschrieben werden kann. So oder so werden in den nächsten Jahren Aspekte höherer Dialogsteuerung – wie sie die Erkennung von Themen darstellt – zunehmend in den Bereich der *interaction robotics* einfließen (müssen), da nur auf diese Weise das Ziel der Konstruktion eines *robot companion* erfüllt werden kann.

Literaturverzeichnis

[Allan u. a. 1999] Allan, J. ; Jin, H. ; Rajman, M. ; Wayne, C. ; Gildea, D. ; Lavrenko, V. ; Hoberman, R. ; Caputo, D.: Topic-based novelty detection / summer workshop at clsp. 1999. – Forschungsbericht

[Allan 2002a] Allan, James: Introduction to Topic Detection. In: Allan, James (Hrsg.): *Topic Detection and Tracking.* Norwell, Massachusetts : Kluwer Academic Publishers, 2002, Kap. 1, S. 1–16

[Allan 2002b] Allan, James (Hrsg.): *Topic Detection and Tracking.* Norwell, Massachusetts : Kluwer Academic Publishers, 2002

[Allan u. a. 2000] Allan, James ; Lavrenko, Victor ; Jin, Hubert: First story detection in TDT is hard. In: *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management.* New York, NY, USA : ACM Press, 2000, S. 374–381

[Allan u. a. 1998] Allan, James ; Papka, Ron ; Lavrenko, Victor: On-Line New Event Detection and Tracking. In: *Research and Development in Information Retrieval*, 1998, S. 37–45

[Altmann 1981] Altmann, G.: Funktionsanalyse in der Linguistik. In: Esser, Jürgen (Hrsg.) ; Hübler, Axel (Hrsg.): *Forms and Functions.* Tübingen, 1981, S. 25–32

[Ando 2000] Ando, Rie K.: Latent Semantic Space: Iterative Scaling Improves Precision of Inter-document Similarity Measurement. In: *Proceedings of the 23rd SIGIR*, 2000, S. 216–223

[Apté u. a. 1994a] Apté, Chidanand ; Damerau, Fred ; Weiss, Sholom M.: Automated Learning of Decision Rules for Text Categorization. In: *ACM Transactions on Information Systems* (1994)

[Apté u. a. 1994b] Apté, Chidanand ; Damerau, Fred ; Weiss, Sholom M.: Toward Language Independent Automated Learning of Text Categorization Models. In: *SIGIR94*, 1994

[Baum 1900] Baum, Lyman F.: *The Wonderful Wizard of Oz.* G. M. Hill, 1900

[Beeferman u. a. 1997] Beeferman, Doug ; Berger, Adam ; Lafferty, John: Text Segmentation Using Exponential Models. In: Cardie, Claire (Hrsg.) ; Weischedel, Ralph (Hrsg.): *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing.* Somerset, New Jersey : Association for Computational Linguistics, 1997, S. 35–46

[Beeferman u. a. 1999] Beeferman, Doug ; Berger, Adam ; Lafferty, John D.: Statistical Models for Text Segmentation. In: *Machine Learning* 34 (1999), Nr. 1-3, S. 177–210

[Ben-Hur u. a. 2001] Ben-Hur, Asa ; Horn, David ; Siegelmann, Hava T. ; Vapnik, Vladimir: Support Vector Clustering. In: *Journal of Machine Learning Research* (2001), Nr. 2, S. 125–137

[Bestgen 2006] Bestgen, Yves: Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001). In: *Computational Linguistics* 32 (2006), March, Nr. 1, S. 5–12

[Black u. a. 1999] Black, A. W. ; Taylor, P. ; Caley, R.: *The Festival Speech Synthesis System – System documentation for Festival Version 1.4.0.* 1999. – `http://www.cstr.ed.ac.uk/projects/festival/manual/`

[Boersma und Weenink 2001] Boersma, Paul ; Weenink, David: PRAAT, a system for doing phonetics by computer. In: *Glot International* 5 (2001), Nr. 9/10, S. 341–345

[Boersma und Weenink 2005] Boersma, Paul ; Weenink, David: *Praat: doing phonetics by computer.* 2005. – (Version 4.3.14) [Computer program] `http://www.praat.org/`

[Bortz und Lienert 2003] Bortz, Jürgen ; Lienert, Gustav A.: *Kurzgefasste Statistik für die Klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben.* 2. Auflage. Berlin : Springer, 2003

[Brindöpke u. a. 1997] Brindöpke, Christel ; Häger, J. ; Johanntokrax, Michaela ; Pahde, Arno ; Schwalbe, Michael ; Wrede, Britta: Darf ich Dich Marvin nennen? Instruktionsdialoge in einem Wizard-of-Oz-Szenario: Szenario-Design und Auswertung / Bielefeld University, SFB360. 1997. – Forschungsbericht

[Bronstein u. a. 2001] Bronstein, I.N. ; Semendjajew, K.A. ; Musiol, G. ; Mühlig, H.: *Taschenbuch der Mathematik.* 5. Auflage. Thun und Frankfurt am Main : Verlag Harry Deutsch, 2001

[Brown und Yule 1983] Brown, Gillian ; Yule, George: *Discourse Analysis.* Cambridge : Cambridge University Press, 1983

[Bußmann 2002] Bussmann, Hadomut: *Lexikon der Sprachwissenschaft.* 3. Auflage. Stuttgart : Alfred Kröner Verlag, 2002

[Cahn und Brennan 1999] Cahn, Janet E. ; Brennan, Susan E.: A Psychological Model of Grounding and Repair in Dialog. In: Brennan, Susan E. (Hrsg.) ; Giboin, Alain (Hrsg.) ; Traum, David (Hrsg.): *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems.* Menlo Park, California : American Association for Artificial Intelligence, 1999, S. 25–33

[Choi 2000] Choi, Freddy: Advances in domain independent linear text segmentation. In: *Proceedings of NAACL-00*, 2000

[Choi u. a. 2001] Choi, Freddy ; Wiemer-Hastings, Peter ; Moore, Johanna: Latent Semantic Analysis for Text Segmentation. In: *Proceedings of 6th EMNLP*, 2001, S. 109–117

[Chomsky 1965] Chomsky, Noam: *Aspects of the Theory of Syntax*. Cambridge, Massachusetts : MIT Press, 1965

[Cieri u. a. 2002] Cieri, Christopher ; Strassel, Stephanie ; Graff, David ; Martey, Nii ; Rennert, Kara ; Liberman, Mark: Corpora for Topic Detection and Tracking. In: Allan, James (Hrsg.): *Topic Detection and Tracking*. Norwell, Massachusetts : Kluwer Academic Publishers, 2002, S. 33–66

[Cimiano u. a. 2004] Cimiano, Phillip ; Hotho, Andreas ; Staab, Steffen: Comparing Conceptual, Partitional and Agglomerative Clustering for Learning Taxonomies from Text. In: *Proceedings of the European Conference on Artificial Intelligence (ECAI'04)*, IOS Press, 2004, S. 435–439

[Clark 1992] Clark, Herbert H. (Hrsg.): *Arenas of Language Use*. University of Chicago Press, 1992

[Coradeschi und Saffiotti 2001] Coradeschi, Silvia ; Saffiotti, Alessandro: Perceptual Anchoring of Symbols for Action. In: *IJCAI*, 2001, S. 407–.416

[Crystal 1995] Crystal, David: *Die Cambridge Enzyklopädie der Sprache*. Frankfurt, New York : Campus Verlag, 1995. – Studienausgabe

[Deerwester u. a. 1990] Deerwester, Scott C. ; Dumais, Susan T. ; Landauer, Thomas K. ; Furnas, George W. ; Harshman, Richard A.: Indexing by Latent Semantic Analysis. In: *Journal of the American Society of Information Science* 41 (1990), Nr. 6, S. 391–407

[Dempster u. a. 1977] Dempster, A. ; Laird, N. ; Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistcal Society* 39 (1977), Nr. B, S. 1–38

[van Dongen 2000] Dongen, Stijn van: *Graph Clustering by Flow Simulation*. Utrecht, University of Utrecht, Dissertation, May 2000. – http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm

[Duda u. a. 2001] Duda, Richard O. ; Hart, Peter E. ; G.Stork, David: *Pattern Classification*. 2. Auflage. John Wiley and Sons, Inc., 2001

[Dutoit u. a. 1996] Dutoit, T. ; Pagel, V. ; Pierret, N. ; Bataille, F. ; Vrecken, O. V. der: The MBROLA project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non Commercial Purposes. In: *Proc. ICSLP '96* Bd. 3. Philadelphia, PA, 1996, S. 1393–1396

[Elmasri und Navathe 2002] Elmasri, Ramez ; Navathe, Shamkant B.: *Grundlagen von Datenbanksystemen*. 3. Auflage. Pearson Studium, September 2002

[Fink 1999] Fink, Gernot A.: Developing HMM-based Recognizers with ESMERALDA. In: Matoušek, Václav (Hrsg.) ; Mautner, Pavel (Hrsg.) ; Ocelíková, Jana (Hrsg.) ; Sojka, Petr (Hrsg.): *Lecture Notes in Artificial Intelligence* Bd. 1692. Berlin Heidelberg : Springer, 1999, S. 229–234

[Fong u. a. 2002] Fong, Terrence ; Nourbakhsh, Illah ; Dautenhahn, Kerstin: A survey of socially interactive robots: concepts, design, and applications / Carnegie Mellon University Robotics Institute. 2002 (CMU-RI-TR-02-29). – Forschungsbericht

[Freund und Schapire 1997] Freund, Y. ; Schapire, R. E.: A decision-theoretic generalisation of on-line learning and application to boosting. In: *Journal of computer and system science* 55 (1997), S. 119–139

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

[Fritsch u. a. 2003] Fritsch, J. ; Kleinehagenbrock, M. ; Lang, S. ; Plötz, T. ; Fink, G. A. ; Sagerer, G.: Multi-Modal Anchoring for Human-Robot-Interaction. In: *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems* 43 (2003), Nr. 2–3, S. 133–147

[Giuliani u. a. 1994] Giuliani, D. ; Omologo, M. ; Svaizer, P.: Talker Localization and Speech Recognition Using a Microphone Array and a Cross-Powerspectrum Phase Analysis. In: *International Conference on Spoken Language Processing (ICSLP) 1994*, September 1994, S. 1243–1246

[Greiff u. a. 2000] Greiff, Warren ; Hurwitz, Laurie ; Merlino, Andrew: MITRE TDT-3 segmentation system. In: *TDT-3 Topic Detection and Tracking Conference*. Gathersburg, MD, February 2000

[Haasch u. a. 2005] Haasch, A. ; Hofemann, N. ; Fritsch, J. ; Sagerer, G.: A Multi-Modal Object Attention System for a Mobile Robot. In: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. Edmonton, Alberta, Canada : IEEE, August 2005, S. 1499–1504

[Haasch u. a. 2004] Haasch, A. ; Hohenner, S. ; Hüwel, S. ; Kleinehagenbrock, M. ; Lang, S. ; Toptsis, I. ; Fink, G. A. ; Fritsch, J. ; Wrede, B. ; Sagerer, G.: BIRON – The Bielefeld Robot Companion. In: Prassler, E. (Hrsg.) ; Lawitzky, G. (Hrsg.) ; Fiorini, P. (Hrsg.) ; Hägele, M. (Hrsg.): *Proc. Int. Workshop on Advances in Service Robotics*. Stuttgart, Germany : Fraunhofer IRB Verlag, May 2004, S. 27–32

[Hajime u. a. 1998] Hajime, Mochizuki ; Takeo, Honda ; Manabu, Okumura: Text segmentation with multiple surface linguistic cues. In: *Proceedings of the 36th annual meeting on Association for Computational Linguistics*. Morristown, NJ, USA : Association for Computational Linguistics, 1998, S. 881–885

[Halliday und Hasan 1976] Halliday, M. A. K. ; Hasan, R.: *Cohesion in English*. London : Longman, 1976

[Hayes u. a. 1990] Hayes, Philip J. ; Anderson, Peggy M. ; Nirenburg, Irene B. ; Schmandt, Linda M.: TCS: A Shell for Content-Based Text Categorization. In: *IEEE Conference on Artificial Intelligence Applications*, 1990

[Hayes und Weinstein 1990] Hayes, Philip J. ; Weinstein, Steven P.: CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories. In: *Second Annual Conference on Innovative Applications of Artificial Intelligence*, 1990

[Hearst 1997] Hearst, Marti A.: TextTiling: Segmenting text into multi-paragraph subtopic passages. In: *Computational Linguistics* 1 (1997), Nr. 23, S. 33–64

[Hearst und Plaunt 1993] Hearst, Marti A. ; Plaunt, Christian: Subtopic structuring for full-length document access. In: *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM Press, 1993, S. 59–68

[Hegel u. a. 2006] Hegel, Frank ; Spexard, Thorsten ; Vogt, Thurid ; Wrede, Britta ; Horstmann, Gernot: Playing a different imitation game: Interaction with an Empathic Android Robot. In: *Proc. 2006 IEEE-RAS International Conference on Humanoid Robots (Humanoids´06)*, IEEE, December 2006

[Hockett 1958] Hockett, C. F.: *A course in modern linguistics*. New York : Macmillian, 1958

[Hofemann u. a. 2004] Hofemann, N. ; Fritsch, J. ; Sagerer, G.: Recognition of Deictic Gestures with Context. In: Rasmussen, C. E. (Hrsg.) ; Bülthoff, H. H. (Hrsg.) ; Giese, M. A. (Hrsg.) ; Schölkopf, B. (Hrsg.): *DAGM04* Bd. 3175. Heidelberg, Germany : Springer-Verlag, 2004, S. 334–341

[Hoffmann 1993] Hoffmann, Ludger: Thema und Rhema in einer funktionalen Grammatik. In: Eisenberg, P. (Hrsg.) ; Klotz, P. (Hrsg.): *Sprache gebrauchen – Sprachwissen erwerben*. Stuttgart : Klett, 1993, S. 135–147

[Hoffmann 2000] Hoffmann, Ludger: Thema, Themenentfaltung, Makrostruktur. In: al., G. Antos/K. B. et (Hrsg.): *Text- und Gesprächslinguistik* Bd. 1. Berlin/New York : de Gruyter, 2000, S. 344–356

[Hofmann 1999] Hofmann, Thomas: Probabilistic Latent Semantic Analysis. In: *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*. Stockholm, 1999

[Hofmann 2001] Hofmann, Thomas: Unsupervised learning by probabilistic latent semantic analysis. In: *Machine Learning* 42 (2001), S. 177–196

[Hohenner u. a. 2003] Hohenner, S. ; Lang, S. ; Kleinehagenbrock, M. ; Fink, G. A. ; Kummert, F.: Multimodale Sprecherlokalisation für Mensch-Roboter-Interaktionen in einer Multi-Personen-Umgebung. In: Kroschel, K. (Hrsg.): *Elektronische Sprachsignalverarbeitung* Bd. 28. Karlsruhe, 2003, S. 162–169

[Hohenner 2005] Hohenner, Sascha: *Automatische Spracherkennung für agierende Systeme*, Universität Bielefeld, Technische Fakultät, Dissertation, 2005

[Hotho u. a. 2005] Hotho, Andreas ; Nürnberger, Andreas ; Paass, Gerhard: A Brief Survey of Text Mining. In: *Zeitschrift fuer Computerlinguistik und Sprachtechnologie (GLDV-Journal for Computational Linguistics and Language Technologie)* 20 (2005), Nr. 1, S. 19–62

[Hüwel und Wrede 2006] Hüwel, Sonja ; Wrede, Britta: Spontaneous Speech Understanding for Robust Multi-Modal Human-Robot Communication. In: *Proceedings of the International Conference on Computational Linguistics (COLING/ACL)*, ACL Press, 2006

[Hüwel u. a. 2006] Hüwel, Sonja ; Wrede, Britta ; Sagerer, Gerhard: Robust speech understanding for multi-modal human-robot communication. In: *Proc. 15th Int. Symposium on Robot and Human Interactive Communication*, IEEE Press, 2006

[Jin u. a. 1999] Jin, H. ; Schwartz, R. ; Sista, S. ; Walls, F.: Topic tracking for radio, tv broadcast and newswire. In: *Proceedings of the DARA Broadcast News Workshop*, Morgan Kauffman Publisher, February 1999, S. 199–224

[Joachims 1997] Joachims, Thorsten: Text Categorization with Support Vector Machines: Learning with Many Relevant Features / Universität Dortmund. 1997. – LS8-Report 23. URL: http://www.cs.cornell.edu/People/tj/publications/joachims_97b.pdf

[Joachims 1998] Joachims, Thorsten: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, Claire (Hrsg.) ; Rouveirol, Céline (Hrsg.): *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Chemnitz, DE : Springer Verlag, Heidelberg, DE, 1998, S. 137–142

[Keenan und Schieffelin 1976] Keenan, E. ; Schieffelin, B.: Topic as a discourse notion: A study of topic in the conversations of children and adults. In: Li, C. (Hrsg.): *Subject and topic*. New York : Academic Press, 1976

[Kim u. a. 2003] Kim, Yu-Seop ; Chang, Jeong-Ho ; Zhang, Byoung-Tak: An Empirical Study on Dimensionality Optimization in Text Mining for Linguistic Knowledge Acquisition. In: Whang, K. Y. (Hrsg.) ; Jeon, J. (Hrsg.) ; Shim, K. (Hrsg.) ; Srivatvava, J. (Hrsg.): *Lecture Notes in Artificial Intelligence* Bd. 2637. 2003, S. 111–116

[Klein und Stutterheim 1992] Klein, Wolfgang ; Stutterheim, Christiane v.: Textstruktur und referentielle Bedeutung. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 86 (1992), Nr. 22, S. 67–92

[Kleinehagenbrock u. a. 2002] Kleinehagenbrock, M. ; Lang, S. ; Fritsch, J. ; Lömker, F. ; Fink, G. A. ; Sagerer, G.: Person Tracking with a Mobile Robot

based on Multi-Modal Anchoring. In: *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*. Berlin, Germany : IEEE, September 2002, S. 423–429

[Kleinehagenbrock 2005] Kleinehagenbrock, Marcus: *Interaktive Verhaltenssteuerung für Robot Companions*, Universität Bielefeld, Technische Fakultät, Dissertation, 2005

[Köhler 1986] Köhler, Reinhard: *Quantitative Linguistics*. Bd. 31: *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum : Studienverlag Dr. N. Brockmeyer, 1986

[Kohonen 1981] Kohonen, Teuvo: Automatic formation of topological maps of patterns in a self-organizing system. In: Oja, E. (Hrsg.) ; Simula, O. (Hrsg.): *Proceedings of 2SCIA, Scand. Conference on Image Analysis*. Helsinki, Finland, 1981, S. 214–220

[Landauer und Dumais 1997] Landauer, Thomas K. ; Dumais, Susan T.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Aquisition, Induction and Representation of Knowledge. In: *Psychological review* 104 (1997), Nr. 2, S. 211–240

[Lane u. a. 2003] Lane, Ian R. ; Kawahara, Tatsuya ; Matsui, Tomoko: Language model switching based on topic detection for dialog speech recognition. In: *Proc. ICASSP* Bd. 1, 2003, S. 616–619

[Lang u. a. 2003] Lang, S. ; Kleinehagenbrock, M. ; Hohenner, S. ; Fritsch, J. ; Fink, G. A. ; Sagerer, G.: Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot. In: *Proc. Int. Conf. on Multimodal Interfaces*. Vancouver, Canada : ACM, November 2003, S. 28–35

[Lang 2005] Lang, Sebastian: *Multimodale Aufmerksamkeitssteuerung für einen mobilen Roboter*, Universität Bielefeld, Technische Fakultät, Dissertation, 2005

[Leopold 2005] Leopold, Edda: On Semantic Spaces. In: *LDV-Forum* 20 (2005), Nr. 1, S. 63–86

[Lewis 1992a] Lewis, David D.: An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In: *Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992, S. 37–50

[Lewis 1992b] Lewis, David D.: Feature Selection and Feature Extraction for Text Categorization. In: *Proceedings of Speech and Natural Language Workshop* Defense Advanced Research Projects Agency (Veranst.), Morgan Kaufmann, February 1992, S. 212–217

[Lewis und Ringuette 1994] Lewis, David D. ; Ringuette, Marc: A Comparison of Two Learning Algorithms for Text Categorization. In: *Symposium on Document Analysis and Information Retrieval*. Las Vegas, NV, April 1994, S. 81–93

[Li 2006] Li, Shuyin: A dialog system for comparative user studies on robot verbal behavior. In: *Proc. 15th Int. Symposium on Robot and Human Interactive Communication.*, IEEE Press, 2006

[Li u. a. 2005] Li, Shuyin ; Haasch, Axel ; Wrede, Britta ; Fritsch, Jannik ; Sagerer, Gerhard: Human-style interaction with a robot for cooperative learning of scene objects. In: *Proc. Int. Conf. on Multimodal Interfaces*. Trento, Italy : ACM Press, 2005, S. 151–158

[Li u. a. 2006] Li, Shuyin ; Wrede, Britta ; Sagerer, Gerhard: 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

[Litman und Passonneau 1995] Litman, Diane J. ; Passonneau, Rebecca J.: Combining Multiple Knowledge Sources for Discourse Segmentation. In: *Meeting of the Association for Computational Linguistics*, 1995, S. 108–115

[Lohninger 2005] Lohninger, H.: *Grundlagen der Statistik*. url: http://www.statistics4u.info/fundstat_germ/cc_distance_meas.html. Januar 2005. – EBook

[Lömker 2004] Lömker, Frank: *Lernen von Objektbenennungen mit visuellen Prozessen*, Universität Bielefeld, Technische Fakultät, Dissertation, 2004

[Lowe 2001] Lowe, Will: Towards a theory of semantic space. In: *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, 2001, S. 576–581

[Luhn 1958] Luhn, H. P.: The automatic creation of literature abstracts. In: *IBM Journal of Research and Development* (1958), April, S. 155–164

[Lund und Burgess 1996] Lund, Kevin ; Burgess, Curt: Producing high-dimensional semantic spaces from lexical co-occurrence. In: *Behavior Research Methods, Instrumentation, and Computers* (1996), S. 203–20

[Maas und Wrede 2006] Maas, Jan F. ; Wrede, Britta: BITT: A Corpus for Topic Tracking Evaluation on Multimodal Human-Robot-Interaction. In: *Proceedings of the international conference on Language and Evaluation (LREC)*. Genoa, Italy, 2006

[Maas u. a. 2006a] Maas, Jan F. ; Spexard, Thorsten ; Fritsch, Jannik ; Wrede, Britta ; Sagerer, Gerhard: BIRON, what's the topic? – A Multi-Modal Topic Tracker for improved Human-Robot Interaction. In: *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, IEEE, September 2006

[Maas u. a. 2006b] Maas, Jan F. ; Wrede, Britta ; Sagerer, Gerhard: Towards a Multimodal Topic Tracking System for a Mobile Robot. In: *Proc. International Conference on Spoken Language Processing (Interspeech/ICSLP)*, September 2006

[Mehler 2001] Mehler, Alexander: *Textbedeutung. Zur prozeduralen Analyse und Repräsentation struktureller Ähnlichkeiten von Texten*. Frankfurt am Main : Verlag Peter Lang GmbH, 2001 (Sprache, Sprechen und Computer)

[Mehler 2004] Mehler, Alexander: Textmining. In: Lobin, Henning (Hrsg.) ; Lemnitzer, Lothar (Hrsg.): *Texttechnologie. Perspektiven und Anwendungen*. Tübingen, 2004, S. 329–352

[Mehler 2006] Mehler, Alexander: Compositionality in Quantitative Semantics. A Theoretical Perspective on Text Mining. In: Mehler, Alexander (Hrsg.) ; Köhler, Reinhard (Hrsg.): *Aspects of Automatic Text Analysis* Bd. 209. Berlin / Heidelberg : Springer, 2006

[Milde u. a. 1997] Milde, J. T. ; Peters, K. ; Strippgen, S.: Situated communication with robots. In: *First Int. Workshop on Human-Computer-Conversation*. 1997

[Mitkov 2002] Mitkov, Ruslan: *Anaphora Resolution*. 1. Auflage. Edinburgh, UK : Longman (Pearson Education), 2002

[Mori 1982] Mori, Masahiro: *The Buddha in the robot*. Boston, MA : Tuttle, 1982

[Morris 1988] Morris, Jane: Lexical Cohesion, the Thesaurus, and the Structure of Text / Computer Systems Research Institute, University of Toronto. Toronto, 1988. – Forschungsbericht. Technical Report CSRI219

[Morris und Hirst 1991] Morris, Jane ; Hirst, Graeme: Lexical cohesion computed by thesaurial relations as an indicator of the structure of text. In: *Computational Linguistics* 17 (1991), Nr. 1, S. 21–48. – ISSN 0891-2017

[van Mulbregt u. a. 1999] Mulbregt, P. van ; Carp, I. ; Gillick, L. ; Lowe, S. ; Yamron, J.: Segmentation of Automatically Transcribed Broadcast News Text. In: *Proceedings of the DARPA Broadcast News Workshop* DARPA (Veranst.), Morgan Kauffman Publishers, February 1999, S. 77–80

[Nakata u. a. 2002] Nakata, Takayuki ; Ikeda, Takahiro ; Ando, Shinichi ; Okumura, Akitoshi: Topic Detection based on Dialogue History. In: *Proceedings of the Workshop on Speech-to-Speech Translation Algorithms and Systems*. Philadelphia, July 2002, S. 9–14

[Ontrup und Ritter 2005] Ontrup, Jörg ; Ritter, Helge: A hierarchically growing hyperbolic self-organizing map for rapid structuring of large data sets. In: *Proceedings of the 5th Workshop on Self-Organizing Maps (WSOM 05)*. Paris, France, Sep 2005. – URL: http://www.techfak.uni-bielefeld.de/ags/ni/publications/media/OntrupRitter2005-AHG.pdf

[Pflüger 2004] Pflüger, Jörg: Konversation, Manipulation, Delegation: Zur Ideengeschichte der Interaktivität. In: Hellige, Hans D. (Hrsg.): *Geschichten der Informatik: Visionen, Paradigmen, Leitmotive*. Berlin, Heidelberg : Springer, 2004, S. 367–410

[Pickering und Garrod 2004] Pickering, Martin J. ; Garrod, Simon: Towards a mechanistic psychology of dialogue. In: *Behavioral and Brain Sciences* 27 (2004), Nr. 2, S. 169–190

[Ponte und Croft 1997] Ponte, J. ; Croft, W.: Text Segmentation by Topic. In: *ECDL '97: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, Morgan Kauffman Publishers, 1997, S. 113–125

[Prassler u. a. 1999] Prassler, E. ; Scholz, J. ; Fiorini, P.: Navigating a Robotic Wheelchair in a Railway Station during Rush Hour. In: *Int. Journal on Robotics Research* 18 (1999), Nr. 7, S. 760–772

[Purver u. a. 2006] Purver, Matthew ; Körding, Konrad P. ; Griffiths, Thomas L. ; Tenenbaum, Joshua B.: Unsupervised Topic Modelling for Multi-Party Spoken Discourse. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia : Association for Computational Linguistics, July 2006, S. 17–24. – URL: http://www.aclweb.org/anthology/P/P06/P06-1003

[Rehder u. a. 1998] Rehder, B. ; Schreiner, M. E. ; Wolfe, M. B. W. ; Laham, D. ; K., Landauer T. ; Kintsch, W.: Using Latent Semantic Analysis to assess knowledge: some technical considerations. In: *Discourse Processes* Bd. 25, 1998, S. 337–354

[Rennie 2004] Rennie, Jason D. M.: *Derivation of the F-measure*. February 2004. – URL: http://people.csail.mit.edu/jrennie/writing

[Rieger 1981] Rieger, Burghard B.: Feasible Fuzzy Semantics. On some problems of how to handle word meaning empirically. In: H. J. Eikmeyer, H. R. (Hrsg.): *Words, Worlds and Contexts*. de Gruyter, 1981, S. 193–209

[Rieger 1989] Rieger, Burghard B.: *Unscharfe Semantik: die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten*. Frankfurt am Main : Verlag Peter Lang GmbH, 1989

[van Rijsbergen 1979] Rijsbergen, C. J. van: *Information Retrieval*. London : Butterworths, 1979

[Sahlgren 2001] Sahlgren, Magnus: Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels. In: *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*. Helsinki, Finland, 2001

[Salton und McGill 1983] Salton, G. ; McGill, M. J.: *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983

[Schmid 1994] Schmid, Helmut: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *International Conference on New Methods in Language Processing*. Manchester, UK, 1994

[Schmid 1995] Schmid, Helmut: Improvements in part-of-speech tagging with an application to German. In: Feldweg (Hrsg.) ; Hinrichs (Hrsg.) ; Feldweg (Hrsg.) ; Hinrichs (Hrsg.): *Lexikon und Text*. 1995, S. 47–50

[Schmidt u. a. 2006] Schmidt, J. ; Kwolek, B. ; Fritsch, J.: Kernel Particle Filter for Real-Time 3D Body Tracking in Monocular Color Images. In: *Proc. of Automatic Face and Gesture Recognition*. Southampton, UK, April 2006, S. 567–572

[Schultz und Liberman 1999] Schultz, J. M. ; Liberman, M.: Topic Detection and Tracking using idf-Weighted Cosine Coefficient. In: *Proceedings of the DARPA Broadcast News Workshop*, 1999, S. 189–192

[Schölkopf und Smola 2002] Schölkopf, Bernhard ; Smola, Alex: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. Cambridge, MA : MIT Press, 2002

[Schütze 1992] Schütze, Hinrich: Dimensions of meaning. In: *Proceedings of Supercomputing '92, Minneapolis.*, 1992, S. 787–796

[Schütze 1998] Schütze, Hinrich: Automatic Word Sense Discrimination. In: *Computational Linguistics* 24 (1998), Nr. 1, S. 97–123

[Seo und Sycara 2004] Seo, Young W. ; Sycara, Katia: Text Clustering for Topic Detection / Robotics Institute, Carnegie Mellon University. January 2004. – Forschungsbericht

[Serafin 2003] Serafin, Riccardo: *Feature Latent Semantic Analysis for dialogue act interpretation*. Chicago, University of Illinois, Diplomarbeit, 2003

[Serafin und Eugenio 2004] Serafin, Riccardo ; Eugenio, Barbara D.: FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification. In: *ACL 2004*, 2004, S. 692–699

[Shriberg u. a. 2000] Shriberg, Elizabeth ; Stolcke, Andreas ; Hakkani-Tür, Dilek ; Tür, Görkhan: Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. In: *Speech Communication* 1–2 (2000), Nr. 32, S. 127–154

[Spexard u. a. 2006] Spexard, Thorsten ; Li, Shuyin ; Wrede, Britta ; Fritsch, Jannik ; Sagerer, Gerhard ; Booij, Olaf ; Zivkovic, Zoran ; Terwijn, Bas ; Kröse, Ben: BIRON, where are you? - Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, IEEE, October 2006

[Spärck Jones 1972] Spärck Jones, Karen: A statistical interpretation of term specificity and its application in retrieval. In: *Journal of Documentation* 28 (1972), S. 11–21

[Stolcke u. a. 1999] Stolcke, Andreas ; Shriberg, Elizabeth ; Hakkani-Tür, Dilek ; Tür, Görkhan ; Rivlin, Ze'ev ; Sönmez, Kemal: Combining Words and Speech Prosody for Automatic Topic Segmentation. In: *Proc. DARPA Broadcast News Workshop*. Herndon, VA, 1999, S. 61–64

[Viola und Jones 2002] Viola, Paul ; Jones, Michael: Robust Real-time Object Detection. In: *International Journal of Computer Vision* 57 (2002), Nr. 2, S. 137–154

[Wada u. a. 2002] Wada, Kazuyoshi ; Shibata, Takanori ; Saito, Tomoko ; Tanie, Kazuo: Analysis of factors that bring mental effects to elderly people in robot assisted activity. In: *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, IEEE Press, 2002, S. 1152–1157

[Walters u. a. 2005] Walters, M. L. ; Dautenhahn, K. ; Koay, K. L. ; Kaouri, C. ; Boekhorst, R. te ; Nehaniv, C. L. ; Werry, I. ; Lee, D.: Close encounters: Spatial distances between people and a robot of mechanistic appearance. In: *Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids2005)*. Tsukuba International Congress Center, Japan, December 2005, S. 450–455

[Wiemer-Hastings 2000] Wiemer-Hastings, P.: Adding syntactic information to LSA. In: *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ : Erlbaum, 2000, S. 989–993

[Wiemer-Hastings und Zipitria 2001] Wiemer-Hastings, P. ; Zipitria, I.: Rules for Syntax, Vectors for Semantics. In: *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Mahwah, NJ : Erlbaum, 2001

[Wrede u. a. 2004] Wrede, S. ; Fritsch, J. ; Bauckhage, C. ; Sagerer, G.: An XML Based Framework for Cognitive Vision Architectures. In: *Proc. Int. Conf. on Pattern Recognition*, 2004, S. 757–760

[Yang 1997] Yang, Yiming: An Evaluation of Statistical Approaches to Text Categorization. In: *Information Retrieval* 1 (1997), Nr. 1/2, S. 69–90

[Youmans 1991] Youmans, Gilbert: A new tool for discourse analysis: the vocabulary management profile. In: *Language* 67 (1991), S. 763–789

[Zipf 1949] Zipf, G. K.: *Human behaviour and the principle of least effort*. Reading, MA : Addison-Wesley, 1949

8 Anhang

8.1 Experimentbogen – Korpuserstellung

Auf der folgenden Seite ist der Experimentbogen abgedruckt, den die Versuchspersonen, die an der Erstellung des BITT-Korpus teilgenommen haben, ausgehändigt bekamen.

Dear participant,

thank you very much for participating in our experiment.
This experiment is part of a test and enhancement phase of our mobile robot BIRON (see picture below). The goal of this development phase is to enable BIRON to show rooms and places to children at the age of about 7 years. In order to do this BIRON itself has to be familiarized beforehand with the room that it has to present. Your task now is to show this room to BIRON so that it can present it to the children afterwards.

At the moment BIRON undergoes a training and development phase and has therefore only limited capabilities. In the experiment it is not able to move because it has to remain attached to the power cord. Also, it can not yet talk and answer your questions. It will, however, follow you by rotating its base and camera without leaving its place. Also, it will show you a friendly face when it has finished processing your utterance. BIRON will try to follow your face with its top camera. Don‘t feel uncomfortable when it tracks only your face, the robot is able to watch the room with its second camera.

Remember thar BIRON has to show the room to 7 years old children himself: You may want to explain the use of items in the room unknown to children to this age.
Tell the experimentator afterwards when you are ready to begin or when you have any more questions. However, if you have questions about BIRON please wait to ask them till after the experiment.

When you talk to BIRON please speak in your normal speaking style but try to speak clearly and not too fast in order to give BIRON some time to understand what you have said.

When you think you have understood the instructions you can give this sheet back to the instructor.

BIRON

8.2 Fragenkatalog

Die folgenden Fragen wurden von den Versuchspersonen nach dem Experiment beantwortet. Multiple-Choice-Antworten sind in runden Klammern hinter den Fragen angeführt:

1. Fields of interest/research
2. How old are you?
3. Do you smoke? (yes/no)
4. Are you male or female? (male/female)
5. Is English your native language? (yes/no)
6. If English is not your native language: How good is your english language proficiency? (excellent/very good/good)
7. Are you experienced in using speech recognition systems? (yes/no)
8. If this is the case: How often do you use these systems? (regulary/often/rarely/almost never)
9. Do you have a certain dialect?
10. How much experience do you have with robots? (never seen/known from media/seen in live/I have already interacted with)
11. What did you like least about BIRON?
12. Which abilities did you miss in the interaction with BIRON?
13. How natural did you find the interaction with BIRON? (very natural/relatively natural/more unnatural/absolutely unnatural)
14. Did you act differently with the robot than with a human concerning speech? (same as human/slightly simpler/quite simpler/absolutely simpler)
15. Did you act differently with the robot than with a human concerning gestures? (more than with a human/same as with a human/less that with a human/none at all)
16. What do you believe did the roboter understand? (same as a human, i.e., everything/most of what I said/a little bit/nothing)
17. How agreeable did you find the task? (very agreeable/rather agreeable/rather not agreeable/not agreeable at all)
18. Did you feel surveyed? (yes, very much/rather surveyed/less surveyed/no, not at all)
19. Would you like to have an introduction for the interaction with BIRON? (yes/no)

20. How much fun did you have interacting with BIRON? (a lot of fun/some fun/rather less fun/no fun at all)

21. Would you like to have a robot at home? (yes, surely/perhaps/rather not/no, not at all)

22. If rather YES: Which abilities should the robot have?

23. If rather NO: Which abilities should the robot have to be accepted at your home?

8.3 Die BITT-DTD

Im Folgenden ist die DTD abgebildet, anhand der der BITT-Korpus annotiert wurde.

```
<?xml version="1.0" encoding="iso-8859-1"?>

<!-- tdt_1.dtd - Diese DTD dient der manuellen Annotation von
in einem Dialog vokommenden Themen (topics). Die Themen sind
benannt und nicht-hierarchisch.
-->

<!-- dialog ist der vorgesehene Root-Knoten -->
<!ELEMENT dialog (utterance|answer|topic)*>
<!ATTLIST dialog
        topic_source CDATA #IMPLIED
        name CDATA  #REQUIRED>

<!-- Ein utterance ist ein Abschnitt der Spracheingabe.
     utterances sind durch Pausen segmentiert. In einem Anwendungssystem
     entstehen sie durch den Spracherkenner.
     syllables ist die Anzahl der Silben in einem utterance-->
<!ELEMENT utterance (lemma,orig)>
<!ATTLIST utterance
        start           CDATA  #REQUIRED
        end             CDATA  #REQUIRED
        ta                    CDATA   #IMPLIED
        tb         CDATA   #IMPLIED
        tc CDATA  #IMPLIED>

<!ELEMENT lemma (#PCDATA|object|objects|noise|hum|quest|german|instructor)*>
<!ELEMENT orig (#PCDATA|object|objects|noise|hum|quest|german|instructor)*>

<!ELEMENT object (candidate*,pos)>
<!ELEMENT objects (group,member*,pos)>
<!ELEMENT member EMPTY>
<!ATTLIST member
     oid CDATA #IMPLIED>
<!ELEMENT group EMPTY>
<!ATTLIST group
     oid CDATA #REQUIRED>

<!-- Ein candidate ist ein moegliches Referenzobjekt. Die ID kann sowohl
     neu als auch schon bekannt sein.
     Fuer den Fall, dass mehrere Referenzobjekte moeglich sind, koennen
     Wahrscheinlichkeiten unter 'prob' angegeben werden. -->
<!ELEMENT candidate EMPTY>
<!ATTLIST candidate
         oid  CDATA  #REQUIRED
```

```
prob CDATA #IMPLIED>

<!ELEMENT pos (#PCDATA|german|quest)*>

<!-- Und nun folgt die Spezifikation der Ereignis-Tags -->
<!ELEMENT hum EMPTY>
<!ATTLIST hum
type CDATA #REQUIRED>

<!ELEMENT noise EMPTY>

<!ELEMENT german EMPTY>
<!ATTLIST german
speech CDATA #REQUIRED>

<!ELEMENT quest EMPTY>
<!ATTLIST quest
speech CDATA #IMPLIED>

<!ELEMENT instructor EMPTY>
<!ATTLIST instructor
speech CDATA #REQUIRED>
```

8.4 Im online-Experiment verwendete Objekte und Themengruppen

Anmerkung: Im Folgenden sind die Objekte und Objektgruppen aufgezählt, die im *online*-Experiment verwendet wurden. Dabei sind sowohl die antizipierten Bezeichner für die Themen, als auch die für die Objekte dargestellt. Objekte, die potentiell mehreren Themen zugeordnet werden können (wie z.B. die Teekanne) wurden der Einfachkeit halber bei nur jeweils einem Thema aufgeführt.

Unterstriche stehen für Leerzeichen, die aufgrund unterschiedlicher Schreibweisen zustande kommen können.

Themengruppe 1:
Themennamen: computer, electronics, technical_devices
Maus: mouse, computer_mouse, computermouse
Tastatur: keyboard, keypad, console
Mousepad: mousepad, pad
Notebook: notebook, laptop, computer
Memorystick: memorystick, memory_stick, usbstick, usb_stick

Themengruppe 2:
Themennamen: sweets, candy, goodies, tuck, food, something_to_eat, somethingtoeat, ailment, edibles, foods, foodstuff, nourishment
Snickers: snickers, chocolate, chocolate_bar, chocolatebar
Keksdose: cookies, box_of_cookies, boxofcookies, biscuits, cookiebox
Bonbon: bonbon, bon_bon, bon-bon, bonbon, candy
Dauerlutscher: lolly, lollypop
Waffeln: wafers

Themengruppe 3:
Themennamen: toys, toy_animals, toyanimals, puppets, soft_toys
Stoffelefant: elephant, toyelephant, toy_elephant
Stoffkrokodil: crocodile, crocodile, toycrocodile, toy_crocodile, alligator
Plüschelch: elk, toyelk, toy_elk, moose toymoose, toy_moose, reindeer, toyreindeer, toy_reindeer
Stoffente: duck, toy_duck, toyduck, bird, toy_bird, toybird
Stoffkuh: cow, toycow, toy_cow

Themengruppe 4:
Themennamen: making_tea, teamaking, preparing_tea, tea, cooking_tea
Becher: mug, cup, beaker, teacup
Teekanne: teapot, pot
Tee: tea, box_of_tea, teabox, boxoftea, packet_of_tea, packetoftea
Stövchen: teapot_warmer, teapotwarmer, stove
Teelöffel: spoon, teaspoon, tea_spoon

8.5 Im online-Experiment verwendete Anleitungsbögen

Dargestellt sind die beiden Anleitungsbögen, die vor dem ersten bzw. zweiten Experimentteil an die Versuchspersonen ausgeteilt wurden.
Anmerkung: Es handelt sich um die Anleitungsbögen für die Hauptgruppe. In dem Anleitungsbogen fur die Kontrollgruppe wurden die Hinweise auf die Möglichkeit, das aktuelle Thema zu erfragen, entfernt.

Vielen Dank für die Teilnahme am Experiment!

Ziel des Experiments ist eine Unterhaltung mit dem Roboter BIRON über verschiedene Objekte. Beachten Sie bitte, dass der Roboter **nur Englisch** spricht und versteht.

1. Teil:

Machen Sie sich erst einmal mit dem Roboter vertraut. Begrüßen Sie den Roboter. Dann können Sie ihn fragen, welche Fähigkeiten er hat und ihm anschließend verschiedene Objekte zeigen. Auf dem Tisch hinter Ihnen befinden sich Objekte, die Sie dafür verwenden können. Legen Sie die Objekte auf die gekennzeichneten Stellen damit BIRON sie sehen kann. Achten Sie bitte darauf, auf der gekennzeichneten Stelle zu stehen.

Nach ein paar Minuten wird der Experimentator Sie unterbrechen und Teil 1 des Experiments beenden.

Wichtig dabei:

1. **Warten Sie bitte nach jedem Satz und machen Sie eine Pause.** Der Roboter braucht manchmal etwas mehr Zeit für die Verarbeitung.
2. **Warten Sie jedes mal auf eine Reaktion des Roboters**, nachdem Sie etwas gesagt haben. Geben Sie ihm Zeit zu reagieren.
3. Bitte sprechen Sie während des Experiments kein Deutsch. Der Roboter kann Deutsch und Englisch nicht unterscheiden.
4. Falls BIRON „I have lost you“ sagen sollte, begrüßen Sie ihn einfach neu, damit er Sie als Kommunikationspartner wiederfinden kann.

BIRON

Abbildung 8.1: Anleitungsbogen vor dem ersten Experimentteil

2. Teil:

Es ist Ihre Aufgabe, dem Roboter thematische Zusammenhänge zwischen Objekten beizubringen.

BIRON versteht in diesem Experimentteil nur folgende Sätze:

- „Hello Biron“
- „This is a ...“
- „These are ...“
- „We talk about ...“
- „What are we talking about?“

Gehen Sie folgendermaßen vor:

1) Begrüßen Sie BIRON
2) Nehmen Sie zwei oder drei Objekte vom Tisch hinter Ihnen. Die Objekte sollen thematisch zusammengehören – das ist es, was BIRON lernen soll.
3) Stellen Sie die Objekte auf die markierten Bereiche auf dem Tisch vor BIRON.
4) Benennen Sie für BIRON die Objekte („this is a...“) und zeigen Sie gleichzeitig auf das benannte Objekt. Zeigen Sie jedes Objekt einzeln und warten Sie, bis BIRON es sich angesehen hat.
5) Stellen Sie alle Objekte wieder zurück und fahren Sie fort. Nehmen Sie auch ruhig Objekte, die Sie BIRON schon vorher gezeigt haben, um sein Wissen zu testen.
6) Sagen Sie BIRON nach Belieben, worüber Sie mit ihm reden („We talk about“).
7) Fragen Sie BIRON nach Belieben, worüber Sie mit ihm reden („What are we talking about“).

Fahren Sie so lange fort, bis der Experimentator Sie nach wenigen Minuten unterbricht.

Viel Spass!

Abbildung 8.2: Anleitungsbogen vor dem zweiten Experimentteil

8.6 Im online-Experiment verwendete Fragebögen

Die folgenden Seiten zeigen die in den Experimenten verwendeten Fragebögen (in verkleinerter Darstellung).

Liebe/r Teilnehmer/in, bitte nehmen Sie sich vor dem nächsten Experimentteil ein paar Minuten Zeit und füllen diesen Fragebogen aus!

User Nummer	Alter	Geschlecht ☐m ☐w	Studiengang; Semesteranzahl/Beruf

1. Ist Ihnen die Kontaktaufnahme mit BIRON leicht gefallen?

Sehr leicht	*Eher leicht*	*Eher schwer*	*Sehr schwer*
☐	☐	☐	☐

2. Wieviel Erfahrung haben Sie im Umgang mit Robotersystemen?

Noch nie gesehen	*Bekannt aus Medien*	*Live gesehen*	*Mal selbst mit interagiert*
☐	☐	☐	☐

3. Wie natürlich fanden Sie die Interaktion mit BIRON?

Sehr natürlich	*Relativ natürlich*	*Eher unnatürlich*	*Völlig unnatürlich*
☐	☐	☐	☐

4. Hatten Sie das Gefühl, der Roboter versteht Sie?

Immer	*Meistens*	*Eher selten*	*Gar nicht*
☐	☐	☐	☐

5. Für wie intelligent halten sie BIRON?

Sehr intelligent	*Intelligent*	*Wenig intelligent*	*Unintelligent*
☐	☐	☐	☐

6. Hätten Sie mit einem Menschen anders interagiert?

a)...was die Sprache betrifft?

Genau wie mit einem Menschen	*Etwas einfacher*	*Sehr vereinfacht*	*Völlig vereinfacht*
☐	☐	☐	☐

b)...oder die Gestik?

Mehr als bei Menschen verwendet	*Genau wie bei Menschen*	*Weniger als bei Menschen*	*Gar keine*
☐	☐	☐	☐

Abbildung 8.3: Fragebogen nach dem erstem Experimentteil, Teil 1

7. Glauben Sie, dass BIRON verstanden hat, was Sie ihm erklärt haben?

So viel wie ein Mensch, alles	*Das meiste*	*Eher wenig*	*Nichts*
☐	☐	☐	☐

8. Glauben Sie, dass BIRON in der Lage war, im Dialog Zusammenhänge zu erkennen?

Ja	*Größtenteils schon*	*Größtenteils nicht*	*Nein*
☐	☐	☐	☐

9. Würden Sie die Interaktion mit BIRON als „intuitiv“ bzw. menschenähnlich beschreiben?

Ja	*Größtenteils schon*	*Größtenteils nicht*	*Nein*
☐	☐	☐	☐

Begründung: .....

10. Fühlten Sie sich in Ihrer Sprache eingeschränkt?

Nein	*Größtenteils nicht*	*Größtenteils schon*	*Ja*
☐	☐	☐	☐

Wenn ja, was haben sie anders gemacht?

.....

11. Wie haben Sie die Aufgabe empfunden?

Angenehm	*Eher angenehm*	*Eher unangenehm*	*Völlig unangenehm*
☐	☐	☐	☐

12.Welche Fähigkeiten von BIRON fanden Sie am besten?

_____

_____

13. Was hat Ihnen an BIRON am wenigsten gefallen?

_____

_____

Abbildung 8.4: Fragebogen nach dem erstem Experimentteil, Teil 2

14. Haben Sie sich beobachtet gefühlt?

Ja, sehr	*Größtenteils schon*	*Größtenteils nicht*	*Nein, gar nicht*
☐	☐	☐	☐

15.Wieviel Spaß hat Ihnen die Interaktion gemacht?

Sehr viel Spaß	*Etwas Spaß*	*Eher keinen Spaß*	*Überhaupt keinen Spaß*
☐	☐	☐	☐

Vielen Dank! Bitte teilen Sie dem/der Experimenator/in mit, wenn Sie fertig sind.

Abbildung 8.5: Fragebogen nach dem erstem Experimentteil, Teil 3

Liebe/r Teilnehmer/in, vielen Dank für Ihre Hilfe! Bitte nehmen Sie sich noch ein paar Minuten Zeit und füllen diesen Fragebogen aus!

User Nummer

1. Empfanden Sie Ihr Treffen mit BIRON als erfolgreich?

Sehr erfolgreich	*Größtenteils*	*Eher wenig*	*Gar nicht erfolgreich*
☐	☐	☐	☐

2. Fanden Sie, dass BIRON seine Aufgaben gut bewältigt hat?

Ja, immer	*meistens*	*manchmal*	*Nein, nie*
☐	☐	☐	☐

3. Hatten Sie das Gefühl, dass der Roboter Sie versteht?

Immer	*Meistens*	*Eher selten*	*Gar nicht*
☐	☐	☐	☐

4. Für wie intelligent halten Sie BIRON?

Sehr intelligent	*Intelligent*	*Wenig intelligent*	*Unintelligent*
☐	☐	☐	☐

5. Wie zufrieden sind Sie mit den Antworten von BIRON?

Sehr zufrieden	*Zufrieden*	*Wenig zufrieden*	*Unzufrieden*
☐	☐	☐	☐

6. Als wie gut beurteilen Sie die Fähigkeit von BIRON, Zusammenhänge zu beurteilen?

Sehr gut	*Gut*	*Eher schlecht*	*Schlecht*
☐	☐	☐	☐

7. Schätzen Sie ein: Wieviel hat BIRON von dem, was Sie ihm erklärt haben, verstanden?

So viel wie ein Mensch, alles	*Das meiste*	*Eher wenig*	*Nichts*
☐	☐	☐	☐

Abbildung 8.6: Fragebogen nach dem zweitem Experimentteil, Teil 1

8. Was hätte BIRON besser machen können, um seine Aufgabe zu bewältigen?

.....

9. Würden Sie die Interaktion mit BIRON als „intuitiv“ bzw. menschenähnlich beschreiben?

Ja	*Größtenteils schon*	*Größtenteils nicht*	*Nein*
☐	☐	☐	☐

Begründung: .....

10. Welche Fähigkeiten von BIRON (wenn überhaupt) fanden Sie am besten?

.....

11. Welche Probleme von BIRON (wenn überhaupt) fanden Sie besonders ärgerlich?

.....

12.Wie haben Sie die Aufgabe empfunden?

Angenehm	*Eher angenehm*	*Eher unangenehm*	*Völlig unangenehm*
☐	☐	☐	☐

13.Wieviel Spaß hat Ihnen die Interaktion gemacht?

Sehr viel Spaß	*Etwas Spaß*	*Eher keinen Spaß*	*Überhaupt keinen Spaß*
☐	☐	☐	☐

14.Hätten Sie gerne einen Roboter zu Hause?

Auf jeden Fall	*Vielleicht*	*Eher nicht*	*Auf gar keinen Fall*
☐	☐	☐	☐

16.a Falls eher JA:
Welche Fähigkeiten sollte der Roboter haben?

_____

_____

16.b Falls eher NEIN:
Welche Fähigkeiten müsste ein Roboter haben, damit Sie ihn bei sich zu Hause akzeptieren würden?

_____

_____

Abbildung 8.7: Fragebogen nach dem zweitem Experimentteil, Teil 2

15.Lassen Sie sich von dem/der Experimentator/in die von BIRON gebildeten Themengruppen zeigen. Als wie gut bewerten Sie sie?

Sehr gut	*Gut*	*Eher schlecht*	*Schlecht*
☐	☐	☐	☐

:-) Vielen Dank! :-)

Abbildung 8.8: Fragebogen nach dem zweitem Experimentteil, Teil 3

Index