3. Methoden

3.1 Wissenschaftstheoretische Grundlegung

Der empirischen Studie sind bestimmte theoretische Grundannahmen vorausgesetzt. Erkenntnistheoretisches Fundament dieser Arbeit ist die Einsicht, dass Theorie und Empirie einander ergänzen und dass Wissenschaft nur möglich ist, wenn Hypothesen und Modelle sowohl theoretisch widerspruchsfrei gedacht als auch durch entsprechende Sprachdaten empirisch belegt werden. Axiomatik sind hierbei die Gesetze der formalen Logik. Auch wenn vorbewusste Aspekte der menschlichen Kognition und subjektive Konzepte in die Theorie miteinbezogen werden, beziehen sich alle Aussagen auf eine per se objektiv fassbare Welt, in der eindeutige Gesetzmäßigkeiten herrschen, auch wenn diese nicht in ihrer Totalität erfassbar sind. Eine wirkliche Synthese zwischen Theorie und Empirie versucht den drohenden Dualismus zwischen intelligibler Welt und sinnlich erfahrbarer Welt zu überbrücken. Diese Herausforderung ist jedoch nicht trivial und stellt nach wie vor ein Schlüsselproblem der Wissenschaftstheorie dar, das hier nicht gelöst werden kann. Eine sinnvolle Ergänzung von Theorie und Empirie wird dennoch angestrebt. Diese Voraussetzungen haben Implikationen für a) die theoretische Struktur der Studie und b) den konkreten Umgang mit Sprachdaten.

Die Korpusanalyse, die in diesem Kapitel genauer expliziert wird, gründet als empirische Studie auf theoretischen Annahmen: die systemtheoretischen Rahmenbedingungen der Interaktion nach Habermas (1993b), die strukturelle Organisation der Gespräche, wie die Konversationsanalyse sie beschreibt, interaktives Alignment als vorbewusster Mechanismus der dialogischen Sprachverarbeitung, logisch-semantische Aspekte der Kohärenz und Kohäsion innerhalb des gegebenen Verweisrahmens sowie die Interpretation und Definition der HCI. Untersuchungsparameter werden entsprechend deduktiv aus theoretischen Modellen abgeleitet. Wo die Struktur sprachlicher Äußerungen im Fokus steht, werden die Untersuchungsparameter aus Grammatikmodellen (deskriptive lexikalische und syntaktische Analyse) deduziert. Eine ideale deduktive Studie gründet auf wenigen Axiomen, ist theoretisch widerspruchsfrei, d.h., folgt den Regeln der Logik und bezieht empirische Daten eng auf theoretische Sätze. Auf diese Weise erzielte Ergebnisse müssen ebenfalls in sich widerspruchsfrei anhand der Gesetze der Logik formuliert werden können. Da die Ergebnisse der Korpusanalyse zur Implementierung verbesserter Subsysteme genutzt werden können sollen, ist deren formale Operationalisierbarkeit von besonderem Interesse. Die auf diese Weise

erzielten Ergebnisse lassen sich in einer formalen Sprache ausdrücken und in ein artifizielles System integrieren. In Kapitel 5.2 werden Überlegungen zur Implementierung dieser Parameter vorgestellt.

Eine induktive Cluster-Analyse aus den Sprachdaten selbst heraus wird für die Studie nicht als sinnvoll erachtet, da in der HHC bereits leistungsfähige theoretische Modelle zur Analyse von Dialogen bestehen. Diese wurden in Kapitel 2 zusammengeführt und sollen in den folgenden Kapiteln auf die HCI übertragen werden.

3.2 Methodischer Zugang

Aus den grundsätzlichen erkenntnistheoretischen Überlegungen leitet sich der methodische Zugang zur empirischen Studie ab. Die in Kapitel 2 dargestellten Ansätze aus den Disziplinen der Konversationsanalyse, der Psycholinguistik, der Forschung zu Kohärenz in Textlinguistik und Pragmatik und der HCI-Forschung zu *Computer-Talk* haben auf theoretischer Ebene zu Hypothesen über Phänomene und Mechanismen der Interaktion mit artifiziellen Agenten geführt, die im Rahmen der Studie empirisch überprüft werden sollen. Da die verschiedenen linguistischen Teildisziplinen traditionell unterschiedliche methodologische Zugänge wählen, musste eine Kombination von Methoden gefunden werden, mit der alle Parameter adäquat untersucht werden konnten. Die vielschichtige Interaktionsform *HCI* kann nur mittels sowohl quantitativer als auch qualitativer Verfahren angemessen analysiert werden (vgl. auch Fischer im Druck: 65). Phänomene, die an einzelnen Wortformen oder Phrasenstrukturen festgemacht werden können, lassen sich gut quantifizieren. Komplexere Konzepte wie Dialogkohärenz oder funktionale Aspekte der Interaktion, können nur durch qualitative Analysen erfasst werden. Hierbei wird ein ethnomethodologischer Ansatz aus der CA gewählt, der zu einem feinkörnigen Verständnis der untersuchten Dialoge unter Einbeziehung des Kontexts führt (vgl. auch Fischer im Druck: 69). Die qualitativen Analysen wurden in einer unabhängigen Untersuchung von einer Hilfskraft gegengeprüft. Da qualitative Analysen aber dennoch die Gefahr bergen, dass ihre Ergebnisse nicht vergleichbar sind, sollen sie durch quantitativ erhobene vergleichbare Werte auf ein sicheres Fundament gestellt werden. Dazu wird ein korpusbasierter Ansatz gewählt.

Es handelt sich um eine Überblicksstudie, mit deren Hilfe ein Gesamtbild der HCI herausgearbeitet werden soll. Die untersuchten Teilkorpora stammen von Interaktionen mit unterschiedlichen Systemen aus den Jahren 2000–2006. Die Studie stützt sich auf authentische Dialoge, die unterschiedliche UserInnen ungesteuert mit unterschiedlich sophistizierten Chatbots und ECAs geführt haben.

Ziel der Untersuchung ist eine umfassende Deskription von Felddaten in Form von HCI-Dialogen auf folgenden Ebenen:

- Lexik
- Syntax
- Sprechakte und Struktur der Dialogoberfläche
- Persistenzstruktur

Die Dialoge werden also hinsichtlich dieser Ebenen polyvalent annotiert, um sprachliche Charakteristika der HCI in ganzer Breite erfassen zu können. Die Annotation erfolgte extrem feinkörnig, um breite Analysen zuzulassen. Sowohl Lemmata und lexikalische Kategorien (POS) als auch syntaktische Kategorien, Sprechakte und konversationsanalytische Kategorien sowie einige Adjazenz-Phänomene bzw. Persistenz als Indikator für Alignment und andere Kohäsionsmittel wurden getaggt, um einen genauen Überblick über die strukturellen Eigenschaften der Dialoge zu bekommen. Außergewöhnlich ist die extrem differenzierte Annotation der Phrasenstruktur gestaffelt nach der Anzahl der Verknüpfungsebenen in einer Phrase. So kann ein genaues Bild der Komplexität von Phrasen in UserInnen- und Systembeiträgen nachgezeichnet und verglichen werden.

Ziel der breit gefächerten Annotation ist es, sich einen vielschichtigen Überblick über die Verteilung unterschiedlicher sprachlicher Phänomene auf deskriptiver Ebene zu verschaffen. Anhand der in Kapitel 2 formulieren Hypothesen sollen zusätzlich Vergleiche angestellt werden in Bezug auf die Wirkung der verschiedenen Systeme auf die beschrieben Aspekte der Sprache ihrer UserInnen. Die ökologisch valide, feinkörnig annotierte Datenbasis bildet eine fundierte Grundlage für die Interpretation der Äußerungen in Bezug auf Tendenzen der syntaktischen Simplifizierung (CT) und für ein tieferes Verständnis für syntaktische Persistenzen vor dem Hintergrund einer Gesamtstruktur. Um Spezifika einer HCI definieren zu können, werden die HCI-Korpora in Relation zu einem Vergleichskorpus gesetzt, das Dialoge unter Menschen enthält.

Der Studie liegen 4 Untersuchungskorpora in Form von Chat-Logs mit rund 150 Interaktionen mit unterschiedlich elaborieren Systemen und ein Vergleichskorpus mit rund 50 Dialogen chat-basierter HHC zu Grunde. Die Korpora umfassen ausschließlich vollständige Dialoge. Die Korpora liegen in tokenisierter und lemmatisierter sowie polyvalent annotierter Form vor gemäß den in Kapitel 2 definierten Untersuchungsparametern, zusätzlich stratifiziert nach Part-Of-Speech, Syntax, Funktion / Sprechakt und Persistenzstruktur. Die Auszählung der Annotationen und die statistische Auswertung der Daten werden in Kapitel 3.3.7 erläutert.

In der HCI-Forschung wurde in der Vergangenheit UserInnen-Verhalten häufig im Experiment untersucht. Vor allem funktionale Aspekte der HCI wurden in experimentellen Settings erforscht (vgl. Fischer 2006). Experimente sind sinnvoll bei der Untersuchung von Einzelphänomenen, eignen sich jedoch nicht für Überblicksstudien. Fischer betont den Vorteil von experimentellen Designs in der HCI-Forschung, da das deterministische System besser kontrollierbar ist als ein menschlicher Stimuli-Geber: „*Finally, the lack of transparency of the system's behaviour is also of great methodological advantage, since it allows the creation of scripted dialogue that is completely comparable across speakers and conditions*" (Fischer 2010: 2350). Doch Systeme, die in diesem Umfang kontrolliert werden, sind nicht mehr vergleichbar mit den Bots, die unsere Alltagswelt in Form von kommerziellen Agenten im WWW usw. bevölkern. Diese komplexen Systeme verfügen nämlich über derart breite Datenbasen, dass es zu einer so großen Variabilität im Output kommt, dass sogar die EntwicklerInnen dieser Systeme nicht mehr vorhersagen können, was ihr jeweiliges System als nächstes ausgeben wird. Also auch wenn Experimente mit Bots besser kontrollierbar sind, fehlt es Ihnen an ökologischer Validität. Sie haben meist einen sehr engen thematischen Fokus und restringierte Funktionen. Das macht sie zu einem zuverlässigen Tool, wenn eng umrissene Mechanismen untersucht werden sollen. Sofern aber die Perspektive der Untersuchung breiter sein soll, ist die Analyse von Felddaten naheliegend. Wenn man UserInnen-Strategien wirklich verstehen möchte, muss man ungesteuerte, authentische und ganze Dialoge untersuchen, wie sie empirisch vorliegen. Experimente, die ein bestimmtes Verhalten evozieren, können Aussagen über dieses Verhalten innerhalb der experimentellen Bedingungen treffen. Die Übertragung ins Feld kann dann erst der zweite Schritt sein und eine genaue Übertragung ist meist nicht möglich. Experimentelle Studien zur UserInnen-Perzeption aus dem Bereich der Usability-Forschung kommen teilweise zu Ergebnissen, die keine Gültigkeit im Feld beanspruchen können. Spätestens wenn die innovative Dialog-Software im Anwendungstest bei den EndverbraucherInnen doch nicht zuverlässig funktioniert, zeigen sich die Schwächen dieser Test-Verfahren.

Im Kontext der HCI-Forschung, wo sich die meisten Studien auf Experimente stützen, zeichnet sich die vorliegende Studie also durch ihre Validität und durch eine besondere Breite an linguistischen Untersuchungsparametern aus. Diese Art der Feldforschung ist im Bereich der HCI selten, weil freie Dialoge mit virtuellen Agenten zu einer Fülle von ungeordneten Sprachdaten führen und breit angelegte Analysen von Dialogtranskripten dadurch aufwändig werden. Der Fokus

solcher linguistischer Korpusstudien ist jedoch weiter und der Geltungsbereich entsprechend größer.

Eine Korpusanalyse ist allerdings nicht das geeignete Verfahren, um den Einfluss einzelner Dialogstrategien des Systems auf das UserInnen-Verhalten zu testen. Auch nach genauer Analyse der Korpora kann kein kausaler Zusammenhang formuliert werden. UserInnen-Reaktionen auf ein modifiziertes Dialog-Design, das in den Korpora beobachtete Phänomene mit einbezieht, können nur im Experiment beobachtet werden, da nur hier die fraglichen Phänomene isoliert manipuliert werden können. Dies wäre eine sich unmittelbar anschließende, wünschenswerte Folgeuntersuchung.

3.3 Aufbau der Korpora

3.3.1 Zusammensetzung der Korpora: Teilkorpora, Korpusgrößen und Inhalte

Die HCI-Korpora liegen als Textdateien und in Tabellenform vor und umfassen in ihrer Gesamtheit 56.218 Tokens. Bei den Forschungskorpora handelt es sich um Logfiles von medial schriftlichen, quasi-synchronen Dialogen zwischen UserInnen und artifiziellen Systemen (132 Dialoge, 45.114 Wortformen). Darüber hinaus liegt das Vergleichskorpus in Form von Protokollen von Mensch-Mensch-Chats (ebenfalls medial schriftlich und quasi-synchron) vor (51 Dialoge, 5.965 Wortformen). Detailliert untersucht wurde also eine Stichprobe von insgesamt 183 Dialogen (51.079 Wortformen).

Die in ihrer ursprünglichen Form wesentlich umfangreicheren Logfiles der Dialoge mit Twispsy, Karlbot und Elbot wurden von der Firma Artificial Solutions unabhängig zu Forschungszwecken zur Verfügung gestellt. Das Vergleichskorpus entstammt dem Dortmunder Chat-Korpus (Beißwenger 2004) und ist unter http://www.chatkorpus.tu-dortmund.de/files/releasehtml/index.html frei verfügbar. Es handelt sich um Protokolle der chat-basierten Bibliotheksauskunft *on demand* der Universitätsbibliothek Dortmund. Die Logfiles des Systems Max (Multimodal Assembly eXpert) wurden von der Arbeitsgruppe *Wissensbasierte Systeme, KI* (Wachsmuth & Kopp) an der Universität Bielefeld zur weiteren Evaluation zur Verfügung gestellt und in dem Umfang für die Analyse genutzt, der nach der Bereinigung des Korpus' möglich war. Das Vergleichskorpus wurde der Online-Version des Dortmunder Chat-Korpus' (Storrer & Beißwenger 2004) entnommen und in vollem Umfang für die Analyse genutzt.

- **Twipsy:** Die im Twipsy-Korpus verarbeiteten Dialog-Protokolle stammen aus dem Jahr 2000 und wurden im Vorfeld und während der Expo in Hannover aufgezeichnet.

Das Stichprobenkorpus aus den Twipsy-Logfiles ist nach dem Elbot-Korpus mit 12562 Tokens bzw. **9750 Wortformen** das zweitgrößte der Systemkorpora.

- **Karlbot:** Die untersuchten Dialoge mit dem virtuellen CEO Karlbot wurden auf der Vertriebsseite von Kiwilogic aufgezeichnet und stammen aus dem Jahr 2002. Das Karlbot-Korpus ist mit 8171 Tokens bzw. **6693 Wortformen** das kleinste der Agentenkorpora, umfasst aber insgesamt 46 Dialoge.
- **Max:** Die vorliegenden Dialogtranskripte stammen von Interaktionen mit den BesucherInnen des HNFs in Paderborn aus den ersten sieben Wochen nach Eröffnung der Ausstellung (15. Januar bis 6. April 2004). In dieser Zeit wurden 2259 Dialoge geführt. Das komplette Max-Korpus umfasst also 4.702.512 Wortformen. Leider konnte nur ein Bruchteil des Gesamtkorpus im Umfang von 11544 Tokens bzw. **8683 Wortformen** in die Analyse einfließen.
- **Elbot:** Das Korpus stammt aus einer Zeit, als Elbot auf freenet.de eingesetzt wurde. Das von Artificial Solutions zur Verfügung gestellte Korpus ist äußerst umfangreich (31.274 Tokens), so dass es nur auszugsweise für die Analyse genutzt werden konnte. Mit 17333 Tokens bzw. **13062 Wortformen** ist es immer noch das größte der untersuchten Korpora.
- **Das Vergleichskorpus (Bib):** Im Dortmunder Chat-Korpus wird das Teilkorpus *„Chat-basierte Bilbliotheksauskunft on demand der Universitätsbibliothek Dortmund"* unter der Nummer 12030000 geführt. Die Mitschnitte stammen aus der Zeit von 2003 bis 2005. Die chat-basierte Bibliotheksauskunft wurde zu diesem Zeitpunkt von der Universitätsbibliothek Bochum als neuer Service eingeführt. Man kann also davon ausgehen, dass die NutzerInnen keine früheren Erfahrungen mit dieser Anwendung gemacht haben, was die Vergleichbarkeit zu den Agentenkorpora erhöht. Das Teilkorpus *„Bibliotheksinformation"* des Dortmunder Chat-Korpus umfasst 200 Dialoge mit insgesamt 21089 Tokens (Wortformen, Satz- und Sonderzeichen sowie Metainformationen im XML-Format). Davon wurden 19850 Tokens von Menschen produziert (Wortformen, Satz- und Sonderzeichen). Das extrahierte Vergleichskorpus umfasst **5067 Wortformen**.

Tabelle 6: Korpusgrößen der Gesamtkorpora

Gesamtkorpus	**Twipsy**	**Karlbot**	**Max**	**Elbot**	**Vergleichskorpus (Bib)**
Tokens (Metadaten + Satz- und Sonderzeichen + WF)	12.562	8.171	11.544	17.333	6.608
Metadaten	988	512	1.434	1.562	643
Wortformen + Satzzeichen	11.574	7.659	10.110	15.771	5.965
Satzzeichen	1.824	966	1.427	2.709	898
Wortformen	9.750	6.693	8.683	13.062	5.067
Dialoge	27	46	39	20	51

Für die Begrenzung der ursprünglichen Korpusgrößen auf die in der Tabelle dargestellten Werte, ist eine Reihe von Gründen anzugeben, die in den entsprechenden folgenden Unterkapiteln näher erläutert wird:

- Bereinigung
- Veränderung der Datenstruktur
- Auswahl einer Stichprobe zur Erleichterung der manuellen Annotation

Bei allen Dialogen des Gesamtkorpus' handelt es sich um 1:1-Chats (vgl. Beißwenger 2007). Mehrparteienchats wurden von Anfang an von der Analyse ausgeschlossen, da dort ganz andere Mechanismen der Sprachhandlungskoordination greifen als im 1:1-Chat. Diese könnten Einfluss haben auf dynamische Faktoren der Interaktion, so dass z. B. Persistenzen aus diesem Grund anders verteilt sein könnten, was zu verfälschten Ergebnissen führen würde.

3.3.1.1 Metadaten: UserInnen-Gruppen und situativer Kontext

Die Korpora in ihrer im Rahmen der Studie verwendeten Form umfassen (ohne Annotationsmarkierung) ausschließlich Sprachdaten. Die einzige verbleibende Metainformation ist die Kategorisierung der Gesprächsbeiträge als UserInnen- oder Systembeitrag. Grundsätzlich konnten keinerlei sekundäre Daten zu Alter, Geschlecht, Herkunft, Bildung und Computerexpertise der UserInnen erhoben werden, da es sich bei den Dialogprotokollen um einfache Logfiles ohne Metadaten handelt. Aus diesem Grund kann im Rahmen der Analyse nicht nach eindeutigen UserInnen-Gruppen stratifiziert werden. In einigen wenigen Dialogen geben UserInnen Informationen zur Person preis. Ob diese immer wahrheitsgemäß sind, bleibt zu bezweifeln. Die Daten genügen also nicht, um Aussagen zu gender- oder altersspezifischem UserInnen-Verhalten zu machen. Mit Bezug auf den Anwendungskontext des jeweiligen Systems lassen sich die UserInnen-Gruppen unter Vorbehalt einschränken. Mit Twipsy interagierten wohl in erster Linie Expo-BesucherInnen, während Karlbot KundInnen und GeschäftspartnerInnen der Firma Kiwilogic gegenüberstanden. Bei den NutzerInnen von Max handelte es sich um BesucherInnen des HNFs. Unter den UserInnen schienen sich viele SchülerInnen zu befinden. Teilweise werden in den Dialogen Angaben zum Alter gemacht, die allgemein auf ein jüngeres Publikum schließen lassen. Außerdem können zahlreiche UserInnen-Äußerungen im Korpus als jugendsprachlich analysiert werden. Die heterogenste NutzerInnen-Gruppe hat Elbot, denn diese rekrutierten sich in der Zeit, als die Logfiles protokolliert wurden, aus KundInnen von freenet.de. Die Chat-TeilnehmerInnen im Vergleichskorpus sind, sowohl Mitglieder als auch MitarbeiterInnen der Universitätsbibliothek Bochum. Auf Seiten der Auskunft chatten BibliothekarInnen sowie studentische

Hilfskräfte. Auf Seiten der Anfragenden chatten Bibliotheksmitglieder, in erster Linie Studierende. Genau wie bei den Agentenkorpora liegen auch hier keine weiteren, persönlichen Daten vor.

Um die unterschiedlichen Teilkorpora besser vergleichen zu können, wurde darauf geachtet, dass die Dialoge alle dem Gesprächstyp *Beratungsgespräch / Informationsanfrage* zugeordnet werden können. Bei Twipsy, Karlbot und Max handelt es sich um so genannte Infobot-Systeme. Das Vergleichskorpus ist im Kontext einer chat-basierten Bibliotheksauskunft aufgezeichnet worden (vgl. Dortmunder Chat-Korpus, Storrer & Beißwenger 2004). Es handelt sich also auch hier um den situativen Kontext eines digitalen Helpdesks, was in enger Verbindung zur Funktion der oben beschriebenen Infobots steht. Das System Elbot kann als Chatterbot im ursprünglichen Sinne verstanden werden. Als Small-Talk-Agent soll er seine UserInnen in erster Linie unterhalten. Die bei der Zusammenstellung des Korpus' ausgewählten Dialoge entstammen jedoch alle einer Phase, als Elbot als Infobot auf freenet.de eingesetzt wurde (Anfang 2006). Insofern kann man auch hier pragmatische Parallelen zu den anderen Agenten-Korpora und dem Vergleichskorpus feststellen. Außerdem muss angemerkt werden, dass die meisten chat-basierten Bots im Netz – sowohl kommerzielle als auch pädagogische Infobots – über Small-Talk-Funktionen verfügen, da sich gezeigt hat, dass von der Technologie begeisterte UserInnen dazu tendieren, die Systeme auf die Probe zu stellen, indem sie Fragen stellen, die den Kontext der vorgegebenen Anwendung überschreiten. Bei der Entwicklung der Systeme wurde daher auch darauf geachtet, dass adäquate Reaktionen auf antizipierbare Small-Talk-Fragen möglich sind. Die Chats im Vergleichskorpus weisen ebenfalls längere Small-Talk-Sequenzen auf, da die MitarbeiterInnen der Bibliothek die Technologie auch zur internen Kommunikation nutzen und teilweise sogar private Themen besprechen.

3.3.2 Trennung der Dialogkorpora in UserInnen- und Systemkorpus

Die Korpora liegen sowohl als dialogische Gesamtkorpora vor als auch als separate Teilkorpora, in denen nur die jeweiligen Gesprächsbeiträge des jeweiligen Systems oder seiner UserInnen enthalten sind. Die dialogischen Gesamtkorpora sind unerlässlich, um funktionale und interaktive Aspekte der Interaktion sinnvoll annotieren zu können. Für die Auszählung wurde mittels eines Java-Skripts eine automatische Trennung der Korpora jeweils in ein UserInnen- und ein System-Korpus vorgenommen, so dass die UserInnen- und die Systemsprache für jedes Interaktionskorpus separat ausgezählt werden konnte. Auch für das

Vergleichskorpus wurde die Trennung Bibliotheksauskunft und Mitglied vorgenommen. Daraus ergibt sich eine Gesamtmenge von 10 Teilkorpora.

Tabelle 7: Die Teilkorpora

Interaktionskorpora	**Teilkorpora**	
Twipsy	Twipsy-UserIn	Twipsy-System
Karlbot	Karlbot-UserIn	Karlbot-System
Elbot	Elbot-UserIn	Elbot-System
Max	Max-UserIn	Max-System
Vergleichskorpus (Bib)	Bib-Mitglied	Bib-Auskunft

Die quantitative Verteilung zeigt bereits auf der Ebene der Korpusgrößen eindeutige Unterschiede sowohl zwischen UserInnen- und Systemkorpora sowie zwischen den HCI-Korpora und dem Vergleichskorpus zur HHC. Die Länge der UserInnen-Beiträge weicht stark von der der System-Turns ab. Diese Differenz hat Einfluss auf die Größen der Subkorpora. Im Vergleichskorpus fallen die Werte nicht so weit auseinander.

Tabelle 8: Korpusgrößen der Teilkorpora

	Twipsy		**Karlbot**		**Max**		**Elbot**		**Bib**	
	System	**UserIn**	**System**	**UserIn**	**System**	**UserIn**	**System**	**UserIn**	**Auskunft**	**Mitglied**
WF	8.468	1.281	6.150	543	7.379	1.303	10.941	2.121	2.885	2.177
Satzzeichen	1.643	181	922	44	1.285	142	2.330	379	503	394
WF und Satzzeichen	10.111	1.462	7.072	587	8.664	1.445	13.271	2.500	3.388	2.571
Anzahl Turns	454	430	195	145	868	500	747	680	297	278
Ø WF pro Turn	18,7	3,0	31,5	3,7	9,0	2,6	14,6	3,1	9,8	7,8
Ø WF pro Dialog	313,6	47,4	133,7	11,8	189,2	33,4	547,1	106,1	56,6	42,7
Ø Turns pro Dialog	16,8	15,9	4,2	3,2	22,3	12,8	37,4	34,0	5,9	5,5

3.3.3 Homogenisierung des Formats der Korpora

Bevor die in den Tabellen dargestellten Teil- und Subkorpora aus den von den Institutionen zur Verfügung gestellten Ausgangskorpora extrahiert werden konnten, mussten einige Veränderungen bezüglich des Formats an den ursprünglichen

Dateien vorgenommen werden. Die Daten lagen nämlich in sehr unterschiedlicher Form vor, so dass zunächst eine vergleichbare Form gefunden werden musste. Die Logfiles von Twipsy, Karlbot und Elbot umfassten im Original von Artificial Solutions mehrere Textdateien. Diese wurden jeweils zu je einem Dokument pro System zusammengeführt. Bei Max war der Arbeitsschritt der Datenzusammenführung nicht mehr nötig, da die Logfiles bereits in Form einer einzigen Datei vorlagen. Die in XML dargestellten Metadaten aus dem Dortmunder Chat-Korpus konnten leider nicht für die weitere Analyse genutzt werden, da zu den HCI-Korpora keine parallelen Meta-Daten vorlagen, und wurden daher nicht in das Vergleichskorpus zur HHC übernommen.

3.3.4 Bereinigung der Korpora (automatisch und manuell)

Die Ausgangskorpora enthielten eine Reihe unterschiedlicher Artefakte, die automatisch mit manueller Nachbereinigung entfernt wurden. Der Bereinigung wurden alle kompletten Ausgangskorpora unterzogen, da vor diesem Schritt nicht klar war, inwiefern dies die Korpusgrößen beeinflussen würde. Denn je nach Menge der zu entfernenden Artefakte schrumpften die Ausgangskorpora um unterschiedliche Faktoren, so dass sie z. T. nach der Bereinigung für die Untersuchung nicht weiter gekürzt werden mussten. Hätte man zuerst gleichgroße Stichproben gezogen und diese dann bereinigt, hätten diese teilweise nicht genug verwertbares Datenmaterial enthalten. Während die Logfiles von Twipsy und Elbot vergleichsweise wenige Artefakte aufwiesen, waren die Logfiles von Karlbot und Max überaus reich an Artefakten. Besonders zwei Artefakte hatten Einfluss auf die Veränderung der Korpusgröße. In den Elbot-Daten fand sich eine lange Sequenz (203.583 Wortformen), die offensichtlich einen mitprotokollierten Systemfehler darstellte. Das System gab hier immer wieder das gesamte Alphabet aus. Diese Sequenz konnte im zweiten Schritt, der Bereinigung per Hand entfernt werden, da es sich um eine zusammenhängende Kette von System-Turns handelte. Problematischer war die Tatsache, dass ein Großteil der Max-Logfiles nur aus Aufforderungen des Systems an potenzielle UserInnen bestand, auf die keine protokollierte Reaktion erfolgte.[93]

93 Im Nixdorf-Museum wird Max ziemlich unauffällig innerhalb der Ausstellung KI in einer Ecke ausgestellt. Da dies nicht der optimale Platz für einen virtuellen Museumsführer ist, macht das System über eine Sprachausgabefunktion auf sich aufmerksam, sobald sich eine Person nähert (vgl. Kapitel 1). Eine Kamera und ein Programm zur Gesichtserkennung liefern die Information, dass Menschen vorbeigehen (vgl. „Flur-Max"). Jede/r vorbeigehende BesucherIn löst so bei Max die Sprachausgabe mit einer

Dieses überflüssige Datenmaterial war über das gesamte Ausgangskorpus verstreut, hoch frequent und in seiner Form alternierend (vgl. Beispiele). Es musste automatisch nach bestimmten Kriterien entfernt werden. Auch in allen anderen Korpora lagen systematisch wiederkehrende Artefakte vor. Daher wurde für jedes Korpus eine so genannte *Black-List* mit zu entfernenden Wortformen, Zeichen oder Phrasen erstellt. Mittels eines Java-Skripts wurden die Korpora so in einem Arbeitsschritt bereinigt und tokenisiert. Folgende Beiträge kamen auf die *Black-List*:

- **Attention-Getting-Tokens des Systems, auf die keine UserInnen-Eingabe folgt**
  (1) Max: hallo!
  (2) Max: spiel mit mir!
  (3) Max: sprich mit mir!
  (4) Max: schon aufgewacht?

- **Links, die nicht syntaktisch in einen Redebeitrag eingebettet sind**

  Max: Wenn Sie weitere Informationen zu den verschiedenen Beiträgen haben möchten, können Sie mich einfach nach einzelnen Ländern fragen.'script language= "JavaScript""!window.opener.parent.location.href=http://www.expo2000.de/deutsch/teilnehmer/tnindex.html;//--"/script'

  Im Gegensatz dazu im Vergleichskorpus:

  Auskunft: Die Bereichsbibliotheken haben abweichende Öffnungszeiten, die Sie unter http://www.ub.uni-dortmund.de/Ueberuns/OeffnungszeitenBB.html nachschauen können.

  Das Problem bestand darin, dass Twipsy ganze Linklisten von bis zu zehn URLs ausgibt, die mit der eigentlichen Dialogführung nur bedingt in Verbindung stehen. Sie gestalten den Dialog multimodal und sind Charakteristikum einer speziellen Art von HCI, was bei der Interpretation der Daten auch beachtet wurde. Da sich die Anzahl der Link-Listen aber durch die ständig gleichen System-Prompts potenzierte, wurden sie zur Vereinfachung der Auszählung aus den Dialogen entfernt. Stattdessen wurde der Platzhalter ENTFERNT in einer separaten Spalte eingefügt, um die Position der ursprünglich vorhandenen Links in die Interpretation der Daten miteinbeziehen zu können.

Aufforderung zum Dialog aus (Bsp.: „Spiel mit mir!", „Hallo!" „Über die Tastatur kannst Du mit mir sprechen.") Allerdings suchen daraufhin nur sehr wenige BesucherInnen wirklich das Gespräch mit Max, so dass das Gros dieser Auforderungen unbeantwortet bleibt und somit nicht in die Analyse miteinbezogen werden kann.

- **Spezielle Artefakte der einzelnen Logfiles**
  Beispiel: Im Karlbot-Korpus fand sich z. B. grundsätzlich der C++-Quellkode für Anführungszeichen im natürlichsprachlichen Text und nach Zitaten (#quotation#).

- **Komplette Dialoge, in denen das System ausschließlich beschimpft wird**
  Hinzu kamen bei Max große Mengen von Dialogen, in denen UserInnen das System ausschließlich beschimpfen (10 %). Dieses UserInnen-Verhalten scheint von Max und der speziellen Situation im Nixdorf-Museum besonders evoziert zu werden. Die Ausstellung wird täglich von Schulklassen besichtigt und die Schüler chatten oft in großen Gruppen mit dem System. Es kann angenommen werden, dass diese Konstellation ein besonders vulgäres Gruppenverhalten begünstigt. Während die UserInnen der anderen untersuchten Systeme jeweils allein zu Hause am heimischen Computer saßen, als sie ihre Eingaben tätigten, versuchten die größtenteils sehr jungen UserInnen von Max offenbar sich gegenseitig in der Gruppe mit ausfallenden Eingaben zu übertreffen. Da dieses Verhalten stark von anderen Aspekten als den vom System vorgegebenen Parametern abhängt, sollte es nicht in die Analyse einfließen. Daher wurden alle Dialoge, die ausschließlich Beschimpfungen enthalten, auf die *Black-List*[94] gesetzt und als ganze Dialoge automatisch entfernt (217.276 Wortformen). Dieser Umstand führte dazu, dass nur ein geringer Anteil vom Max-Korpus für die Analyse genutzt werden konnte. Das Beispiel illustriert das Ausmaß des Problems.

  (1) UserIn: du bist ein idiot
  (2) Max: he das hab ich ja noch nie gehoert.
  (3) UserIn: du bist ein idiot
  (4) Max: du bist ja richtig originell.
  (5) UserIn: du bist ein idiot
  (6) Max: Erzaehl mir doch mal etwas anderes.
  (7) UserIn: du bist ein idiot
  (8) Max: Du musst mir doch nicht alles doppelt und dreifach sagen.
  (9) UserIn: du bist ein idiot
  (10) Max: Wenn du das nochmal sagst weiss ich es auswendig.
  (11) UserIn: du bist ein idiot
  (12) Max: Du wiederholst dich.
  (13) UserIn: du bist ein idiot
  (14) Max: Ich kann auch wieder das gleiche sagen: Du wiederholst dich.
  (15) UserIn: du bist ein idiot
  (16) Max: Erzaehl mir mal was das ich noch nicht weiss.
  (17) UserIn: du bist ein idiot

94 Dialoge wurden entfernt, wenn zwischen der im Log automatisch generierten Anfangs- und Endmarkierung eines Dialogs jeder UserInnen-Beitrag ein Schimpfwort-Token aus einer anhand eines entsprechenden Online-Lexikons erstellten Wortliste enthielt.

> An diesem Beispiel werden zwar die hinterlegten Antwortmuster des Systems offenbar. Für die Evaluation der Systemfunktionen ist es also nicht uninteressant. Zur Analyse des UserInnen-Verhaltens trägt der Dialog aber nur insofern bei, dass an ihm besonders deutlich wird, dass einige BesucherInnen des HNFs das System testen wollten. Dialoge mit längeren Beschimpfungssequenzen, die aber nicht ausschließlich aus Beschimpfungen bestanden, verblieben selbstverständlich im Korpus.

Das Vergleichskorpus umfasste, wie oben bereits erwähnt, mehrere XML-Dateien. Aus diesen Dateien wurden die reinen UserInnen- und Systembeiträge (ohne Metainformationen) mittels eines weiteren Java-Skripts extrahiert und die XML-Struktur wurde zu Gunsten der Vergleichbarkeit mit den anderen Korpora aufgegeben.

Der automatische Bereinigungsprozess wurde begleitet vom manuellen Bereinigen einzelner Stichproben, um die Zuverlässigkeit der Bereinigungsskripte zu überprüfen, um diese ggf. zu modifizieren. Die automatisierten Prozeduren führten zu sehr unterschiedlichen Veränderungen der Korpusgrößen. Vor allem das Max-Ausgangskorpus war davon betroffen, da es in seinem ursprünglichen Zustand in erster Linie aus für die angestrebte Art der Analyse nicht verwertbaren Daten bestand. Nach dem ersten Arbeitsschritt (automatische Bereinigung) ergab sich folgende Verteilung an aussortiertem Datenmaterial:

Tabelle 9: Anteile Black-List

	Isolierte Dialog-anfänge	**„UserIn“ „System“**	**URLs**	**Quellcode-Artefakte**	***Flaming*-Dialoge**	**XML-Code**	**Ausschuss (gesamt)**
Twipsy	76 Tokens		498 Tokens				574 Tokens
Karlbot	70.000 Tokens	5.500 Tokens	9.000 Tokens	203.583 Tokens			288.146 Tokens
Elbot	4.460 Tokens						4.460 Tokens
Max	4.472.344 Tokens		1.348 Tokens		217.276 Tokens		4.690.968 Tokens
Bib			1.440 Tokens			3.582 Tokens	5.022 Tokens

Eine Nachbereinigung per Hand erfolgte danach wie eingangs beschrieben. In einem zweiten Schritt wurden die Korpora manuell nachbereinigt, um alle Artefakte, die automatisch nicht aufgefunden werden konnten, zu eliminieren. Durch *Black-List* und manuelle Nachbereinigung veränderten sich die Korpusgrößen wie folgt:

Tabelle 10: Größe der Korpora vor und nach der Bereinigung

Korpus	Korpusumfang (gesamt) vorher	Korpusumfang (gesamt) nachher	Ausschuss in Prozent
Twipsy	114.471 Tokens	113.897 Tokens	5,0 %
Karlbot	395.853 Tokens	107.707 Tokens	72,8 %
Elbot	31.274 Tokens	26.814 Tokens	14,7 %
Max	4.702.512 Tokens	11.544 Tokens	99,7 %
Vergleichskorpus	11.630 Tokens	6.608 Tokens	43,2 %

3.3.5 Auswahl der Stichproben

Aus den bereinigten Korpora wurden z. T. Stichproben extrahiert, um die Gesamtkorpusgröße noch einmal zu reduzieren und so zu einem Analysekorpus zu kommen, das von einer Einzelperson getaggt werden konnte. Das Max-Korpus und das Vergleichskorpus hatten bereits nach der Bereinigung nur noch einen Umfang von 11.544 bzw. 6.608 Tokens und mussten folglich nicht weiter gekürzt werden. Überschaubare Untersuchungskorpora mit vollständigen Dialogen waren übriggeblieben, die im nächsten Schritt als Zielgröße für die Stichproben aus den anderen Korpora dienten. Bei Twipsy, Karlbot und Elbot wurde eine Auswahl getroffen, um vergleichbar große Teilkorpora zu erhalten. Ebenfalls in Analogie zum Max-Korpus wurden folgende Unterordner der gesamten Logfiles ausgewählt: Twipsys erste Woche auf der Expo-Website, Karlbots erste Woche auf der Kiwilogic-Homepage und Elbots erste Woche auf freenet.de. Dabei wurde darauf geachtet, dass nur ganze Dialoge in die Korpora einfließen sollten. Daraus resultierte für die Auswahl der Stichproben ein Konflikt zwischen der Anzahl der Dialoge und Anzahl der Wortformen. Waren z. B. im Karlbot-Korpus die Dialoge im Durchschnitt verhältnismäßig kurz, so wurde eine Grenze von ca. 10.000 Wortformen (46 Dialoge) als Gesamtumfang des Korpus' angestrebt. Waren die Dialoge hingegen im Durchschnitt länger (vgl. Elbot), so wurde das Korpus auf ein Minimum von 20 Dialogen beschränkt (ca. 13.000 Wortformen).

3.3.6 Aufbereitung und Annotation der Korpora

3.3.6.1 Automatische Aufbereitung (Tokenisieren, Lemmatisieren, POS-Tagging) und manuelle Nachbearbeitung

Nach einem ersten Test der Bereinigungsskripts wurden die anfallenden Arbeitsschritte mittels eines Java-Skrips (java.class) zusammengefasst. Bereinigen, Tokenisieren, Lemmatisieren und die Annotation nach lexikalischen Kategorien

erfolgte in einem Schritt. Es wurde so tokenisiert, dass in den Arbeitskorpora je eine Wortform (auch Satz- und Sonderzeichen) in je eine Zeile notiert wurde (vgl. Lemnitzer & Zinsmeister 2006). Für die Zuweisung der Lemmata und die Annotation nach Wortformen (Part-of-Speech-Tagging, POS-Tagging) wurde der probabilistische Tree-Tagger der Universität Stuttgart (Schmidt 1994, 1995)[95] in das Skript einbezogen. Für das Deutsche gilt dieser bis heute als leistungsstärkster POS-Tagger für die Annotation der Lemmata und Wortarten. Er liefert auch für CMC-Daten bislang die besten Ergebnisse. Mit den HCI-Dialogen konnte er umgehen mit einer Fehlerquote zwischen 8 % und 10 % je nach Korpus, die per Hand nachannotiert wurde.

Die manuelle Nachannotation umfasste vom Standard abweichende Formen, die der Tagger nicht zuordnen konnte. Außerdem wurden syntaktische Kategorien, Dialogstrukturen und Sprechakte sowie Persistenzen per Hand annotiert.

3.3.6.2 Analysekategorien für das polyvalente Tagging per Hand

Die Stichprobenkorpora haben mit einer realen Größe von 56.218 Tokens eine Gesamtgröße, die für ein manuelles, polyvalentes Taggig durch eine Einzelperson hoch angesetzt ist. Um die in Kapitel 2 dargestellten Untersuchungsparameter quantifizierbar zu machen, ist manuelles Tagging jedoch immer dann die einzige Lösung, wenn das fragliche Phänomen zu komplex ist, um präzise definiert und vom Tagger zuverlässig erkannt werden kann.

Die Korpora wurden wie eingangs erwähnt auf unterschiedlichen linguistischen Ebenen annotiert, um einen möglichst breiten Zugang zu HCI-Dialogen zu ermöglichen. Auf diese Art werden die Korpora in Bezug auf ihre strukturellen Eigenschaften quantitativ vergleichbar und Koinzidenzen können ggf. beobachtet werden. Strukturelle Aspekte können unabhängig von semantischen Aspekten betrachtet werden, so dass auch Dialoge mit thematisch unterschiedlichem Fokus verglichen werden können.

Um streng theoriegeleitet vorgehen zu können, wurden nicht nur einzelne kritische Parameter getaggt, sondern alle Wortarten (POS), alle Phrasenstrukturen (Syntax), alle Sprechakte bzw. Dialogfunktionen (DAMSL) sowie persistente Lexeme und Strukturen als Indikatoren für Alignment. Aus der Kombination dieser Faktoren können sowohl Hinweise auf CT abgeleitet werden als auch auf die Übertragung von Strukturen aus der HHC. Die Gliederung der Annotationskategorien wurde für das Tagging nach linguistischen Ebenen (Lexik, Syntax, Kommunikation) und nicht nach Forschungsfragen (vgl. Hypothesen

95 www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Zugriff 15.06.2016).

in Kapitel 2) vorgenommen, um eine hohe Flexibilität für die Auswertung zu gewährleisten und die annotierten Korpora später für andere Forschungsinteressen weiterhin nutzen zu können.

Auf der Basis einer Evaluation der Frequenz einzelner Strukturmerkmale kann eine tentative Interpretation der Charakteristika einer UserInnen-Sprache zwischen zwei Polen vorgenommen werden:

a) in Form von Merkmalen für eine (vorbewusste) Attribuierung anthropomorpher Eigenschaften an das System durch die Übertragung von erlernten Verhaltensweisen aus der HHC (Alignment, Kohäsion, Höflichkeit)
b) in Form von Merkmalen bewusster UserInnen-Strategien (CT als strukturelles oder funktionales Register)

Außerdem kann die Leistungsfähigkeit der unterschiedlich ausgereiften Systeme anhand der strukturellen Untersuchungsparameter verglichen werden (Gesprächsabbrüche, Störungen, Kohäsionsmittel, Quasi-Persistenz des Systems).

Die Annotation der syntaktischen Kategorien richtet sich nach dem deskriptiven Grammatikmodell nach Eisenberg (1999) und wurde für diese Studie entwickelt. Komplexe Phrasen können auf diese Art feinkörnig annotiert werden, so dass die Komplexität dieser Phrasen für die Untersuchung genau abgebildet werden kann (vgl. CT: Simplifizierung vs. Komplexität).

Die Sprechakte wurden nach einer überarbeiteten Version der konversationsanalytischen Annotationskonvention DAMSL (Dialogue-Act-Markup-Language, Allen & Core 1997) annotiert. Da Besonderheiten der HCI in SWBD-DAMSL (Jurafsky, Schriberg & Biasca 1997) nicht dargestellt werden können, wurde das SWBD-DAMSL-Tag-Set hinsichtlich dieser speziellen Parameter erweitert.

Um dynamische Aspekte semantisch und strukturell nicht restringierter UserInnen-Beiträge erfassen zu können, wurden sowohl lexikalische als auch syntaktische Persistenzen annotiert; d.h. es wurde gekennzeichnet, wo UserInnen Lemmata / Lexeme und / oder syntaktische Konstruktionen des Systems übernehmen und in ihren eigenen Beitrag integrieren (vgl. Alignment Pickering & Garrod 2004, Branigan et al. 2000, Szmrecsanyi 2005). Die Annotation nach Sprechakten liefert auch Zusatzinformationen zu funktionalen Aspekten von Persistenzen in den Korpora.

Die Annotationkonvention für Persistenzen wurde eigens für diese Analyse entwickelt und stützt sich theoretisch auf die Ansätze von Pickering und Garrod (2004) zu Alignment als Lower-Level-Priming, von Branigan et al. (2000) zu syntaktischem Alignment, von Gries (2005) zur korpusbasierten Analyse von Persistenzen und von Szmrecsanyi (2005) zur Alpha- und Alpha-Persistenz.

Die Untersuchungsparameter zur Textkohäsion (vgl. Kapitel 2.4) lassen sich unterschiedlichen linguistischen Ebenen zuordnen und finden sich somit im Tag-Set für die jeweilige Ebene (Konnektiva in POS, Adjazenzellipsen in Syntax usw.). Inkohärente und quasi-inkohärente Turns wurden in SWBD-DAMSL auf der Ebene der Dialogfunktionen getaggt.

Die tokenisierten Korpora wurden in Excel-Tabellen übertragen, so dass die Annotationen in unterschiedlichen Spalten polyvalent hinzugefügt werden konnten. Einer Wortform können so mehrere Annotationen nach unterschiedlichen Kategorien zugeordnet werden. Durch die Zuordnung der Annotationskategorien zu unterschiedlichen linguistischen Ebenen markieren die Tags immer zwei Informationen: das Einzelphänomen und die Ebene. Für die Interpretation von komplexen sprachlichen Phänomenen, die sich auf unterschiedlichen Ebenen manifestieren, ist ein solches Vorgehen unerlässlich.

3.3.6.2.1 Annotationskonvention

Jeder Wortform sind ein Lemma und eine lexikalische Kategorie zugeordnet. Die lexikalischen Kategorien werden nach der Annotationskovention des Stuttgart-Tübingen-Taggers, *STTS*[96], als Tags in Großbuchstaben ohne Klammern dargestellt. Syntaktische Kategorien werden an der letzten an der fraglichen Struktur beteiligten Wortform in spitzen Klammern und Minuskeln annotiert, Sprechakte am Turn-Ende klein und in eckigen Klammern. Bei persistenten Strukturen wird sowohl an der Ausgangsform im FPP als auch an der angepassten Form im SPP annotiert. Die Annotation erfolgt in eckigen Klammern und Majuskeln. Lexikalische Persistenz wird dabei direkt an der jeweiligen Wortform annotiert, syntaktische Persistenz entsprechend an der letzten Wortform der fraglichen Struktur. Zwischen Alpha- und Beta-Persistenz wird unterschieden.

- Lexikalische Kategorien: GROSS (nach STTS)
- Syntaktische Kategorien: <klein> (nach Eisenberg 1999)
- Sprechakte: [klein] (nach SWBD-DAMSL modifiziert für das Deutsche und die HCI)
- Persistenzkategorien: [GROSS] (nach Pickering und Garrod 2004, Branigan et al. 2000, Gries 2005, Szmrecsanyi 2005)

Um die UserInnen-Tags von den System-Tags auch unabhängig vom Text unterscheiden zu können, werden alle System-Tags zusätzlich mit einem Ableitungsstrich (‘) versehen. Diese Konvention erweist sich bei automatischen

96 www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (Zugriff 18.06.2016).

Operationen als praktisch, da Tags über einfache Suchfunktionen aus dem Korpus herausgefiltert und gleichzeitig eindeutig zugeordnet werden können.

3.3.6.2.2 Lexik

Zur Annotation der lexikalischen Kategorien wurde dem Tagger entsprechend das *Stuttgart-Tübingen-Tag-Set* (STTS) verwendet. STTS ist ein Tag-Set von 54 Tags für die Annotation deutscher Texte mit Part-of-Speech-Lables (POS). Richtlinien zur Annotation sind als Postscript und als dvi-Dokument online frei zugänglich.

Grob unterscheidet STTS zwischen Nomina, Verben, Adjektiven, Adverben, Pronomina, Konjunktionen, Partikeln und Interjektionen. Besonders hilfreich für wissenschaftliche Korpusanalysen ist die feinkörnige Unterscheidung der Subklassen innerhalb größerer Wortkategorien. So gliedert sich das Tag-Set im verbalen Paradigma in z. B. in 12 Unterkategorien auf (VVFIN, VVIMP, VVINF, VVIZU, VVPP, VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP). Die Präfixe der Tags zeigen dabei die übergeordnete Kategorie an, während die Gliederungsebenen sich nach rechts immer weiter ausdifferenzieren. Auf der Grundlage einer syntaktischen Analyse können so auch unterschiedliche syntaktische Funktionen in die Bestimmung der lexikalischen Kategorien mit einfließen. So wird bei der Analyse der Adjektive z. B. unterschieden zwischen ADJA und ADJD, also zwischen Adjektiven im attributiven Gebrauch und solchen im adverbialen oder prädikativen Gebrauch. Satzzeichen werden separat annotiert und mit $ gekennzeichnet.

Bei der automatischen Annotation in STTS wird eine lexikalische Kategorie nicht berücksichtigt. Es handelt sich um die schwer zu definierende Klasse der Partikeln, die in STTS den Adverben subsumiert werden. Diskurspartikeln und Modalpartikeln sind für die Analyse von großem Interesse, da ihr Fehlen in der Vergangenheit als Indikator für CT gewertet wurde (vgl. Krause & Hitzenberger 1992). Eine Kennzeichnung dieser Partikelklassen wurde nachträglich manuell durchgeführt.

Die Annotation der Wortarten ist darüber hinaus besonders relevant für Überlegungen zur lexikalischen Vielfalt oder Simplifizierung im Diskurs um CT sowie außerdem für die Analyse von Kohäsionsmitteln an der Dialogoberfläche.

3.3.6.2.3 Syntax

Der Analyse der syntaktischen Kategorien in den Agentenkorpora und im Vergleichskorpus liegt die deskriptive Grammatik nach Peter Eisenberg zu Grunde. Eisenberg liefert ein differenziertes System zur Beschreibung des Deutschen. Der

deskriptive Ansatz erwies sich für die Korpusstudie aus verschieden Gründen als besonders praktikabel. Zunächst ist eine Annotation der Sache nach grundsätzlich deskriptiv, so dass ein deskriptives Grammatikmodell als Leitfaden zur Annotation linguistischer Korpora immanent schlüssig ist.

Einzelnen Phrasen im Korpus können als Annotationen nur einzelne syntaktische Kategorien zugeordnet werden und kein komplexes System von Ableitungsregeln. Zu diesem Zweck wurde ein Instrumentarium gesucht, das Kategorien einzeln und unabhängig von Relationen darstellen kann. Eisenberg setzt eine mehrfach verzweigte Baumstruktur an. Im Baum sind auf Ebene der ersten Geschwisterknoten Nominalgruppen und Präpositionalgruppen mit Satzgliedfunktion je einzeln und unverbunden einander nebengeordnet. Vor dem Hintergrund einer solchen Satzstruktur können einzelne Phrasenstrukturen auch als separate Tags verwendet werden, was einer einfachen Annotation zu Gute kommt.

Darüber hinaus kann Eisenberg nach seiner Notationskonvention strukturelle Aspekte der Sprache weitgehend unabhängig von lexikalischen Aspekten darstellen. Dieser Umstand erleichtert eine Trennung der Ebenen beim polyvalenten Tagging. Selbstverständlich ist diese strikte Trennung ein Konstrukt. Die Grenzen zwischen Struktur, Inhalt und Funktion von sprachlichen Äußerungen verschwimmen und jede Trennung ist modellhaft.[97]

NGr / PGr: Von besonderer Bedeutung für die Studie sind unter anderem komplexe Nominal- und Präpositionalgruppen (vgl. Kapitel 4.4). Mittels einer für diese Korpusanalyse entwickelten Nomenklatur kann die Binnenstruktur der jeweiligen Phrase in einem Annotations-Tag erfasst werden. Dabei steht <ngr> für Nominalgruppe und <pgr> für Präpositionalgruppe. Nach dem Prinzip der Ableitung einer Summenformel aus einer Strukturformel in der Chemie wird die Anzahl der Attribute in Form von Nominalphrasen, Präpositionalphrasen oder Attributsätzen im Phrasenstrukturbaum durch Zahlen wiedergegeben. <ngr3> bezöge sich also auf eine Nominalgruppe mit drei phrasalen Attributen. Die Anzahl der adjektivischen Attribute wird zusätzlich gekennzeichnet durch <adja> für adjektivisches Attribut plus die Anzahl der Adjektive in numerischer Form. Präpositionalgruppen können entweder mit einem einfachen Nominal <pgrn> oder mit einer Nominalgruppe <pgrngr> stehen. Im letzteren Fall kann dann die

97 Diesem Umstand ist es gezollt, dass Ansätze zur Construction Grammar (z. B. Stefanowitsch & Fischer 2008) nicht in die vorliegende Korpusanalyse mit einfließen, obwohl sie sprachliche Strukturen auf kognitive Konzepte beziehen und sich somit gut zur Analyse von interaktivem Alignment im Dialog eignen müssten.

Anzahl weiterer Attribute entsprechend der oben dargestellten Struktur komplexer Nominalgruppen angegeben werden. Die Annotation der NGr erfolgt am Kernnominal, die der PGr an der Präposition als Regens. Phrasenkoordination wird zusätzlich getaggt.

ADJGr / PtGr / ADVGr: Auch komplexe Adjektiv-, Partizip- oder Adverbgruppen werden nach der Anzahl ihrer modifizierenden Elemente nummeriert. Sofern sie in Nominalgruppen integriert sind, werden sie trotzdem separat annotiert.

VGr: Als Verbgruppe (VGr) fasst Eisenberg ausschließlich Verbalkonstruktionen mit Modalverb und Vollverb. Konstruktionen mit Auxiliar fallen nicht in diese Kategorie. Im Korpus werden diese am Modalverb als <vgr> annotiert.

Satzwertige Strukturen und Ellipsen: Hinzu kommen die unterschiedlichen Satz- und Ellipsentypen, Infinitivgruppen und Korrelatsätze. Die Annotation erfolgt an der letzten Wortform des Syntagmas. Satzkoordination von V2-Sätzen wird zusätzlich getaggt.

Tabelle 11: Tag-Set Syntax (mit Beispielen aus den Korpora)

Syntaktische Kategorie	Tag	Anmerkung	Beispiel aus Korpus
Hauptsatz	**<hs>**	Finites Verb + obligatorische Ergänzungen	***Ich kann dir Verschiedenes erklären.***
Kopulasatz	**<ks>**	Kopulaverb + Subjekt (+ Prädikativ)	***Ich bin Max.***
Komplementsatz	**<kompls>**	Subjekt oder Objekt mit Satzstatus	*Schön**, dass wir darüber gesprochen haben.***
Adverbialsatz	**<advs>**	Adverbiale Ergänzung mit Satzstatus	*Du kannst mich unterbrechen, **wenn ich was wiederholen soll**.*
Attributsatz	**<attrs>**	Attribut zu Nominal mit Satzstatus	*Ich bin eine künstliche Person, **die sprechen und gestikulieren kann**.*
Adjazenzellipse	**<ella>**	Elliptisches SPP, das sich mit Hilfe des FPPs auf Vollform zurückführen lässt (nach Busler & Schlobinski 1997)	*Max: Heißt das jetzt ja oder nein?* *UserIn: **Ja!***

Syntaktische Kategorie	Tag	Anmerkung	Beispiel aus Korpus
Koordinationsellipse	**<ellko>**	Elliptischer zweiter Teil einer Koordination (nach Eisenberg 1999)	*Ich bin Pfälzer **und stolz darauf**.*
Restklasse andere Ellipsentypen	**<ell>**	z. B. Handlungsellipsen	***Jetzt aber weiter mit deinem Namen.***
Nominalgruppe (NGr)	**<ngr(+X)>**	Kernnominal + Artikel + attributive Ergänzungen Art und Anzahl der Ergänzungen werden an das Tag angehängt: z. B. <ngr2adja1> für eine NGr mit zwei postnominalen und einem pränominalen Attribut	***Ein anderes Exemplar der Spezies***
Präpositionalgruppe (PGr)	**<pgr(+X)>**	Kernnominal + Präposition als Regens Attribute zum Kennominal werden am Tag annotiert (vgl. NGr): z. B. <pgrngr2>	***in Heinz Nixdorfs Leben***
Adjektivgruppe (AdjGr)	**<adjgr(+X)>**	Adjektiv + Ergänzungen (andere Adjektive, Adverben, Partikeln, Negation) Wie bei NGr, wird die Anzahl der Attribute im Tag vermerkt: z. B. <adjgr2>	***künstlich intelligent***
Adverbgruppe (AdvGr)	**<advgr(+X)>**	Adjektiv + Ergänzungen (Adjektive, Adverben, Partikeln, Negation) Wie bei NGr, wird die Anzahl der Attribute im Tag vermerkt: z. B. <advgr2>	***gar nicht mehr***
Partizipgruppe (PtGr)	**<ptgr>**	Adjektiv + Ergänzungen (Adjektive, Adverben, Partikeln, Negation)	*touristisch **genutzt***

Syntaktische Kategorie	Tag	Anmerkung	Beispiel aus Korpus
Verbgruppe (VGr)	**<vgr>**	Vollverb + Modalverb	*Roboter **können** Menschen nicht **ersetzen**.*
Infinitivgruppe (IGr)	**<igr>**	Infinitiv mit zu oder um zu	***Um** eine Antwort **geben zu können***
Satzkoordination	**<kos>**	Nebenordnung gleichwertiger Strukturen mit Satzstatus (V2)	*Ich wollte Ihre Frage beantworten, **aber** mein Antwortalgorithmus ist gerade abgestürzt.*
Phrasenkoordination	**<kop>**	Nebenordnung gleichwertiger Phrasen ohne Satzstatus	*ein brennender See **und** ein Wald*
Anaphorische Referenz[98]	**<ana>**	Rückbezug auf früheren Turn	***Das** geht dich einen feuchten Kehricht an.*
Kataphorische Referenz	**<kata>**	Vorverweis auf späteren Text	***So** ist heute das Wetter in Bonn: 24,2°C, Sonnenschein*
Satzkorrelat	**<korrs>**	Platzhalter + NS	*Ich werde **dafür** plädieren, **dass die Roboter ihr eigenes Land zugewiesen bekommen.***
Korrelat zu IGr	**<korri>**	Platzhalter + IGr	***Es** ist wirklich schön, **Besuch von einem Menschen zu bekommen.***

Eine präzise Übersicht über die syntaktischen Strukturen ist wichtig a) für den Vergleich zwischen UserInnen-Sprache und Systemsprache sowie b) für den (diachronen) Vergleich zwischen den unterschiedlichen Systemen als auch c) den Vergleich zwischen HCI und HHC im Vergleichskorpus.

Die Annotation der Syntax bildet die syntaktischen Strukturen der Dialogbeiträge detailliert ab. Syntaktische Relationen wurden jedoch nicht getaggt. Grund für diese Fokussierung auf Struktur ist die übergeordnete Fragestellung, ob sich CT durch eine geringere syntaktische Komplexität auszeichne (Krause

98 Bei anaphorischer und kataphorischer Referenz handelt es sich um textlinguistische Kategorien. Sie wurden aus Gründen der Übersichtlichkeit in dieses Tag-Set mitaufgenommen.

& Hitzenberger 1992). Syntaktische Komplexität zu definieren, ist nicht trivial. Sie setzt sich aus mehr Parametern zusammen als aus einer hohen Anzahl komplexer Phrasen und subordinierter Sätze. Syntaktische Relationen und Topologie dürfen eigentlich nicht vernachlässigt werden. Ferner manifestiert sich die Komplexität nicht allein in der Menge der komplexen Elemente, sondern auch in der Beschaffenheit von Verweisrelationen. Alles zusammen ist mit korpuslinguistischen Mitteln schwer greifbar. Die Beschränkung der Annotation auf syntaktische Strukturen wurde aus arbeitsökonomischen Gründen vorgenommen. Eine präzisere Analyse der syntaktischen Komplexität in HCI und CMC wäre wünschenswert. Im Zusammenspiel mit den anderen Untersuchungskategorien bietet die Annotation der syntaktischen Strukturen allerdings bereits genügend Anhaltspunkte für die Interpretation der UserInnen- und System-Beiträge.

3.3.6.2.4 Sprechakte und Dialogstruktur (SWBD-DAMSL)

Um Erkenntnisse aus der Konversationsanalyse auf die Korpora anwenden zu können, ist es wichtig, über ein angemessenes Tag-Set für Dialogstrukturen zu verfügen. Persistenzen und mögliche Aspekte eines Computer-Talks sollen im Kontext des Dialogverlaufs betrachtet werden können. So ist zum Beispiel von besonderem Interesse, bei welchen Sprechakten vermehrt Alignment auftritt und wie diese Sprechakte in den übrigen Dialogverlauf eingebettet sind. Wie verteilen sich Persistenzen auf Frage-Antwort-Sequenzen und welche Gesprächsbeiträge liegen beim System und welche bei der Userin? Zeichnen sich antizipierbare Muster ab? Mit quantitativen Methoden kann diesen Fragen nur bedingt nachgegangen werden. Eine qualitative Untersuchung der relevanten Dialogsequenzen bleibt unerlässlich. Um jedoch so viele Informationen wie möglich, quantitativ erheben zu können, wurden die HCI-Korpora und das Vergleichskorpus nach einer Variante der Annotationskonvention SWBD-DAMSL (SWichBoarD-Dialog-Act-Markup-in-Several-Layers, Jurafsky, Shriberg & Biasca 1997) getaggt.[99]

DAMSL *„defines a set of primitive communicative actions that can be used to analyze dialogs"* (Allen & Core 1997: 1). Ziel war die Entwicklung eines Werkzeugs zur automatischen Annotation von Sprechakten und Dialogstrukturen im Dialogverlauf für die englische Sprache. *„The main purpose of our label set is to label these Switchboard conversations for training stochastic discourse grammars so*

99 Bei SWBD-DAMSL handelt es sich um eine Weiterentwicklung von Allens und Cores DAMSL, das 1997 von der Multiparty-Discourse-Group im Rahmen der DRI-DRI-Meetings entwickelt wurde.

as to build better Language Models (LM) for Automatic Speech Recognition (ASR) of Switchboard" (Jurafsky, Shriberg & Biasca 1997: 1). DAMSL und SWBD-DAMSL sind im Bereich der englischsprachigen Gesprächsforschung weit verbreitet. Große Korpora, wie das TRAINS-Korpus oder das SWITCHBOARD-Korpus wurden mit Hilfe dieser Tag-Sets annotiert. Eine zuverlässige automatische Version für das Deutsche existiert z. Z. leider noch nicht. Bei Sprechaktparsern für das Deutsche muss man mit Fehlerquoten von mindestens 40 % rechnen. Sofern die Sprachdaten stark von der Standardsprache abweichen, werden noch weniger Sprechakte richtig erkannt. Um den Ansprüchen der vorliegenden Studie gerecht zu werden, war aus diesen Gründen die Verwendung einer automatischen Sprechaktannotation ausgeschlossen. Daher fiel die Entscheidung auf die Annotation der oberflächlichen Dialogstrukturen nach einer um besondere Phänomene der HCI erweiterten Variante von SWBD-DAMSL per Hand.

SWBD-DAMSL weist einige Vorteile für die Annotation der Sprechakte und der Dialogstruktur auf. Die Korpora sollen nach strukturellen Kriterien evaluiert werden. Übergeordnete Pläne und Ziele von UserInnen spielen bei diesem Forschungsanliegen eine untergeordnete Rolle. Vielmehr ist eine breite und detaillierte Beschreibung der sprachlichen Performanz auf unterschiedlichen Ebenen erwünscht, um daraus antizipierbare Muster der HCI abzuleiten. SWBD-DAMSL bezieht sich auf eine sog. *Shallow Discourse Structure* (Jurafsky, Shiberg & Biasca 1997: 1), die aussagenlogische, formalpragmatische Parameter außer Betracht lässt. Diese Perspektive auf Dialogstrukturen ergänzt die Annotation auf den Ebenen der Lexik und der Syntax, denen ein deskriptives Modell zu Grunde liegt.

Das SWBD-DAMSL-Tag-Set wurde anhand von Transkripten telefonbasierter Mensch-Mensch-Kommunikation erstellt und beinhaltet einige Tags, die für medial schriftliche Interaktion nicht relevant sind. Einige Annotationskategorien nach SWBD-DAMSL können für die chat-basierte HCI vernachlässigt werden (z. B. *Self-Talk*). „*For any particular project, we would expect that the annotation scheme would be refined to provide further detail on phenomena of interest*" (Allen & Core 1997: 2). Allgemein wurde das Tag-Set aber für aufgabenorientierte Zwei-Parteien-Dialoge entwickelt (vgl. Allen & Core 1997: 2), so dass es sich auf die Interaktion zwischen UserIn und System gut übertragen lässt.

Sprechakte werden so annotiert, dass an ihnen z. B. vermerkt wird, ob es sich um eine *Forward Function* (FPP) oder um eine *Backward Function* (SPP) handelt. Adjazenzstrukturen können auf diese Art und Weise indirekt mitannotiert und im Korpus leichter wieder aufgefunden werden. Aus der Abfolge der Sprechakte lassen sich in einem zweiten Schritt typische Muster ableiten, die u. U. Aufschluss geben können über Charakteristika eines CTs auf der Ebene der Sprechakte.

Das SWBD-DAMSL-Tag-Set[100] umfasst 60 Basistags, die teilweise auch kombiniert werden können. Nach Abzug der Tags, die sich allein auf medial mündliche Kommunikation beziehen oder bereits von anderen Tag-Set abgedeckt werden (Syntax, Persistenz), blieben für das vorliegende Tag-Set 37 Sprechakt-Tags. Diese wurden wiederum um 6 zusätzliche Tags für medial schriftliche HCI erweitert, so dass für die Studie ein Tag-Set von 43 Sprechakt-Tags vorliegt. Die Tabelle zeigt das komplette, modifizierte und vereinfachte Tag-Set.

Tabelle 12: Modifiziertes Tag-Set SWBD-DAMSL-HCI

Kategorie	Tag	Kommentar	Beispiel
statemant-non-opinion	**[sd]**	einfache Aussage, der ein Wahrheitsgehalt zugeordnet werden kann	*Ich habe eine Stahlplatte im Kopf.*
statement-opinion	**[sv]**	persönliche Meinung, subjekiv	*Das klingt spannend.*
yes-no-question	**[qy]**	Ja-Nein-Frage	*Kennst du „Herr der Ringe"?*
w-question	**[qw]**	W-Frage	*Warum müssen Sie immer solche Fragen stellen?*
open-question	**[qo]**	offene Frage, lässt eine nicht restringierte Menge von Antworten zu	*Und nun?*
or-question	**[qr]**	Oder-Frage, nur zwei Antworten sind möglich	*Komisch im Sinne von merkwürdig oder im Sinne von amüsant?*
declarative-question	**[^d]**[101]	Deklarativ-Frage	*Und es fallen dann keine Gebühren an?*
tag-question	**[^g]**	**Tag-Frage, im Deutschen selten**	*Aber du bist perfekt, was?*
action-directive	**[ad]**		*Schieß los!*
offer	**[co]**		*Ich kann sehr gerne nachschauen lassen.*
commit	**[cc]**		*Ich melde mich wieder mal!*

100 Das originale SWBD-DAMSL-Tag-Set findet sich unter: www.stanford.edu/~jurafsky/ws97/manual.august1.html (Zugriff 20.05.2016).

101 Bei den mit dem Symbol ^ gekennzeichneten Tags handelt es sich um Annotationen, die in Verbindung mit einer zweiten Annotationskategorie verwendet werden können, um Zusatzinformationen zu enkodieren.

Kategorie	Tag	Kommentar	Beispiel
command	**[com]**		*Mehr Info*
conventional-opening	**[fp]**		*Hallo, Elbot!*
conventional-closing	**[fc]**		*Auf Wiedersehen.*
non-conventional closing	**[fnc]**	**z. B. Verabschiedung fehlt völlig, Dialog wird einfach abgebrochen**	*...*
explicit-performative	**[fx]**		*Ein frohes, neues Jahr wünsche ich Dir!*
exclamation	**[fe]**	**Medial schriftlich realisiert durch Onomatopoesie, Iteration, Großbuchstaben, Akronyme**	*Haha!,* *Haaaha!,* *Ich SCHREIE,* *LOL*
flaming	**[fl]**		*Grins nich so kackfresse!*
thanking	**[ft]**		*Vielen Dank!*
you're welcome	**[fw]**		*Gern geschehen!*
apology	**[fa]**		*Tut mir leid!*
test	**[fts]**	**UserInnen-Eingaben, mit denen das System getestet werden soll (z. B. Nonsense, oder stark kontextabhängige Äußerungen)**	*Ziegensauger?*
please	**[fpl]**		*Schickst Du mir bitte einmal eine Internetseite?*
accept	**[aa]**		*o.k.*
maybe	**[am]**		*Vielleicht…*
reject	**[ar]**		*Nein, danke!*
signal-non-understanding	**[br]**		*Ich befürchte, ich habe nicht genügend Daten, um intelligent zu antworten.*
acknowledge	**[bh]**		*Ach so.*
repeat-phrase	**[^m]**	**Auto-Repetition, die im psycholinguistischen Diskurs bei menschlichen ProbandInnen als Self-Alignment gedeutet wird**	*Max: Ich wünsch Dir noch einen schönen Tag, Ipke!* *UserIn: Bye-bye* *Max:* ***Ich wünsch Dir noch einen schönen Tag!***

Kategorie	Tag	Kommentar	Beispiel
completion	**[^2]**	**B vervollständigt die Äußerung von A bzw. schließt sie als Koordinationsellipse an.**	*Elbot: Sie sollten die Frage mal einem Talk-Show-Moderator stellen.* *UserIn: oder einem General.*
summarize / reformulate	**[bf]**		*UserIn: Kennst Du Günther?* *Elbot: Günther könnte ein Roboter sein.* *UserIn: Ich meine Günther Jauch.*
appreciation	**[ba]**		*Prima!*
sympathy	**[by]**		*Du bist echt süß!*
downplayer	**[bd]**		*UserIn: Tut mir leid* *Twipsy:* ***Ist schon in Ordnung***
yes answers	**[ny]**		*Ja!*
no answers	**[nn]**		*Nein!*
affirmative non-yes answers	**[na]**		*Eben!*
negative non-no answers	**[ng]**		*Ist es nicht!*
other answers	**[no]**		*soso*
incoherent answer	**[ni]**		*UserIn: Kennst Du Star Treck?* *Elbot: Eigentlich habe ich gar kein richtiges Lieblingstier.*
quasi-coherent answer	**[niq]**	**vorformulierte Ausweichantworten,** ***Default-Antworten*** **(wirken oft eigenartig und unerwartet, aber nicht völlig inkohärent)**	*UserIn: Normal ist das nicht!* *Elbot: Dieses Thema ist etwas traumatisch für mich.*
quoted material	**[^q]**		*Erst nach der Show heißt es* ***„Film ab!"***

Die in der Tabelle grau hinterlegten Kategorien wurden als Tags zur Annotation von HCI hinzugefügt und sollen hier genauer beschrieben weren.

Command [com]: Bestimmte UserInnen-Typen tendieren dazu, einfache, sprachlich stark verkürzte Befehle an das System einzugeben. Dabei handelt es

sich meist um einfache Infinitive („öffnen", „zeigen"), Imperative („stopp", „halt") oder freistehende NGrs („mehr Info"). Pragmatisch entsprechen diese Sprechakte eher der Eingabe eines Befehls in Programmiersprache oder über eine bestimmte Tastenkombination (z. B. Weiterklicken über *Enter* oder *Stgr + C* für „kopieren") als natürlichsprachlicher Kommunikation. In der HHC sind solche Sprachhandlungen äußerst selten und wenn sie auf der Performanzebene beobachtet werden können, unterscheiden sie sich doch konzeptionell von Befehlen in der HCI. Krause, Hitzenberger und Womser-Hacker fanden in ihren Studien Anfang der 1990er Jahre zahlreiche Belege für sprachliche Handlungen im Sinne von [com] und klassifizierten sie als typisch für CT (vgl. z. B. Krause & Hitzenberger 1992). Fischer (2006) kann Belege für [com] nur noch beim UserInnen-Typus *Non-PlayerIn* finden (vgl. Fischer 2006: 121–129). Bei [com] handelt es sich um eine Forward-Looking-Function.

Non-Conversational-Closing [fnc]: [fnc] musste eingeführt werden, um alle Dialogbeendigungen zu kennzeichnen, die nicht mit einer wechselseitigen Verabschiedung schließen. In der HHC kommen solche Dialogabbrüche selten vor, da Gesprächsrahmen als hochgradig konventionalisiert angesehen werden können (vgl. Kapitel 2.4; Brinker & Sager 1989). In der HCI hingegen werden Dialoge oftmals von UserInnen abgebrochen aus Gründen, die in Kapitel 4 näher diskutiert werden sollen. Eine Entsprechung zu [fnc] bei der Gesprächseröffnung im Sinne eines Non-Conversational-Openings ist denkbar, wurde aber zur Annotation der Korpora nicht benötigt, da beim Dialog-Design aller ins Korpus aufgenommenen Agenten darauf geachtet wurde, dass die Systeme den Dialog eröffnen und kurz darstellen, welche Funktion sie erfüllen, z. B. „Hallo, ich bin Max. Über die Tastatur kannst Du mit mir sprechen". [fnc] ist eine Forward-Looking-Function und fällt unter die Kategorie *Communication-Management*.

Flaming [fl]: Alonzo und Aiken definieren *Flaming* als Begriff aus dem Kontext der internet-basierten Kommunikation folgendermaßen: „*the term generally requires hostile intentions characterised by words of profanity, obsecenity, and insults that inflict harm to a person or an organisation resulting from uninhibited behavior*" (Alonzo & Aiken 2004: 205). *Flaming* ist ein notorisches Problem der HCI. UserInnen scheinen aus unterschiedlichen Gründen dazu zu tendieren, Systeme zu beschimpfen (vgl. Diskussion zum Max-Korpus, s. Kapitel 4.4), sei es, um das System zu testen, sei es, weil keinerlei Konsequenzen drohen, oder sei es, um sich in einer Gruppe von UserInnen zu produzieren. Im Gegensatz zur HHC, in der Beschimpfungen die Ausnahme darstellen, gehören Sie in den untersuchten HCI-Korpora zur Tagesordnung. Das Tag-Set musste also entsprechend erweitert werden. Bei [fl] handelt es sich um eine Forward-Looking-Function.

Test [fts]: UserInnen versuchen die Möglichkeiten des Systems auf unterschiedliche Arten zu erproben. So werden in der chat-basierten HCI z. B. sinnlose Tastenkombinationen oder ungewöhnliche Wörter eingegeben, um die Reaktion des Systems zu testen. In der HHC oder der CMC wäre ein solches Verhalten nur unter sehr speziellen Bedingungen denkbar. Als initiative Sprachhandlung, durch die ja gerade eine Systemreaktion provoziert werden soll, ist [fts] selbstverständlich zu den *Forward-Looking-Functions* zu zählen.

Incoherent [ni] und quasi-coherent answer [niq]: In der HHC sind ausschließlich solche Gesprächsbeiträge akzeptabel, die logisch kohärent an den Vorgänger-Turn anschließen, semantische Kontiguität aufweisen und dem situativen Kontext angemessen sind (vgl. Kapitel 2.4). Systembeiträge folgen leider nicht immer in dieser Form kohärent auf UserInnen-Eingaben. Grund für solche Störungen des Dialogverlaufs sind Probleme des Parsings, die in unterschiedlicher Form auftreten können und mit denen Systeme unterschiedlich umgehen. Kann einer Eingabe bei einfachen Chatbot-Systemen z. B. kein Muster zugeordnet werden, greifen unterschiedliche Mechanismen.

1. Die UserInnen-Eingabe wird fälschlicherweise in einer Form interpretiert, die von der Userin nicht intendiert war. Das Parsing-Problem wird nicht erkannt und es wird eine Antwort generiert, die zwar kohärent anschließt an die Fehlinterpretation, nicht aber an die eigentliche Eingabe der Userin. Hier sind zwei Szenarien möglich.

    a) In den meisten Fällen sind solche System-Turns sowohl in der Tiefenstruktur als auch an der Oberfläche inkohärent zum UserInnen-Turn; d. h., es besteht weder semantische Kontiguität, noch logische Kohärenz oder thematische Progression zum UserInnen-Turn noch finden sich sprachliche Marker für Kohäsion im ausgegebenen Text, die einen Bezug zum UserInnen-Turn erkennen lassen. Beiträge, die also sowohl in Bezug auf Kohärenz als auch auf Kohäsion nicht an den vorhergehenden Turn des Gegenübers anschließen, werden als inkohärente Antworten annotiert. Solche Turns sind in der HHC kaum vorstellbar, da sie völlig aus der Progression des Dialogs herausfallen und im gegebenen Kontext unsinnig wirken.
    b) In einigen Fällen weist die logisch inkohärente Systemantwort an der Oberfläche Repetitionen auf, die den Eindruck erwecken, sie stünden in direkter Relation zur UserInnen-Eingabe. Stattdessen handelt es sich um Artefakte des Parsing-Fehlers (z. B. Fehler in der semantischen Interpretation einer Wortform, aber Repetition derselben). Solche System-Turns werden als quasi-kohärente Antworten annotiert, da sie vermeintliche Kohäsionsmarker aufweisen, logisch aber nicht kohärent anschließen.

2. Wird ein Parsing-Fehler dagegen vom System direkt erkannt, greifen unterschiedliche Kompensationsmechanismen:

   a) Direkte Rückfragen an die Userin, die zum Einleiten von Repair als kohärente Reaktion auf die Störung interpretiert werden können. Sie werden als entsprechender Fragentyp annotiert. Die Rückmeldung an die Userin, dass der Beitrag nicht „verstanden" wurde und nach Möglichkeit umformuliert werden soll, ist die kohärenteste, aber auch gleichzeitig die riskanteste für den weiteren Dialogverlauf. Systeme, die ständig ausgeben: „Das verstehe ich leider nicht" stoßen bei UserInnen auf geringe Akzeptanz. Dialoge werden früher abgebrochen, wenn Systeme ausschließlich auf vollständig erkannte Muster reagieren.
   b) Default-Antworten, die Gemeinplätze enthalten und in den meisten Dialogsituationen vergleichsweiche akzeptabel erscheinen, werden ausgegeben, um die Störung zu kaschieren. *„Default or safety-net answers are most commonly selected and will therefore be seen to repeat over a series of conversations"* (Roberts 2006: 3). Verlegenheitsantworten machen also bei einfachen Chatbots den Löwenanteil der Systemausgabe aus. Manchmal fügen sie sich zufällig nahtlos in den Dialog und meist sind sie so allgemein gehalten, dass sie auf jeden Vorgänger-Turn passen.

      „One solution is to write meta-answers so general that they fit anything and, if we're lucky, amuse at the same time" (Roberts 2006: 3).

      (1) Elbot: I am glad you are willing to speak to me so openly about such matters. Now, what was it you said?
      (2) Elbot: I suppose if I could respond to that I'd pass the Turing test.

      Je nach Akzeptabilität in der konkreten Dialogsequenz werden diese als kohärent (ohne Annotation), quasi-kohärent (Kohäsionsmarker vorhanden, aber keine semantische Kontiguität) oder inkohärent (keine Kohäsionsmarker, keine semantische Kontiguität) annotiert.

      Oft wirken diese Beiträge jedoch deplaziert und stoßen bei UserInnen nicht auf große Akzeptanz. Ähnlich wie Menschen, die auf Gemeinplätze auszuweichen pflegen, evozieren Bots mit solchen Beiträgen oft aggressives UserInnen-Verhalten.

      (1) Elbot: Vielleicht liegt das an mehr Erfahrung.
      (2) UserIn: vielleicht auch an mangelnder zahl von bits und bytes
      (3) Elbot: Das Leben als Mensch soll ja nicht so einfach sein [niq]
      (4) UserIn: floskelarsch
      (Elbot 2963–2996)

Bei teilweise analysierten UserInnen-Eingaben (Bsp.: In einem längeren Turn konnte nur ein Schlüsselwort im Lexikon gefunden werden.) kann eine Verlegenheitsantwort (*default answer*) generiert werden, die einzelne Schlüsselwörter aufgreift, aber aussagenlogisch in gar keinen Zusammenhang zum Vorgänger-Turn gesetzt werden kann. Trotzdem wird das gleiche Vokabular verwendet und es entsteht eine schwer greifbare Quasi-Kohärenz oder Oberflächenkohärenz [niq].

(1) UserIn: Ich liebe gekochte Katzen.
(2) Elbot: Eigentlich habe ich gar kein richtiges Lieblingstier.
(Elbot 3656–3668)

Mit [ni] und [niq] wird kein Sprechakt bezeichnet, sondern eine Relation zum Kontext. Für die HHC wäre ein solches Tagging nicht im Sinne von SWBD-DAMSL. In der HCI sind jedoch die einzigen Backward-Looking-Functions, die das System (pragmatisch) vollziehen kann, streng genommen: Parsing (+ regelbasierte Ausgabe) vs. Nicht-Parsing (+ Default-Ausgabe). Aus diesem Grund können [ni] und [niq] als Backward-Looking-Functions für die HCI ins Tag-Set aufgenommen werden.

c) Default-Antworten mit Themawechsel, um ein völlig neues Thema einzuführen oder zu einem übergeordneten Dialogziel zurückzukehren (*Goal-Directedness*). Hierbei handelt es sich um eine Spezialform von 2b) und einen Kniff der Dialog-DesignerInnen. An die eigentliche Default-Antwort anschließend wird ein neues Thema eingeführt, das den Dialogverlauf wieder in einen Bereich des Dialog-Designs zurückführt, der sicherer geparst werden kann. Außerdem kann das System auf diese Weise immer wieder zu seinem ursprünglichen Dialogziel zurückkommen (z. B. Exponate erklären, Ticket verkaufen usw.). Typische Beispiele für zielorientierte Themenwechsel nach gescheitertem Parsing im Sinne eines „*goal-directed behaviours*" (Cassell 2002) finden sich in großer Zahl in den Agenten-Korpora – wie z. B. bei Twipsy:

(1) Twipsy: Jetzt sind Sie platt, was?
(2) User: Wie eine Scholle
(3) Twipsy: Habe ich Ihnen eigentlich schon erzählt, dass die EXPO 2000 nicht nur in Hannover, sondern weltweit stattfindet?
(Twipsy 12277–12308)

Die annotierten Sprechakte und Dialogstrukturen stellen die Basis für verschiedene Untersuchungsbereiche dar. Merkmale der Dialogstruktur nach SWBD-DAMSL fließen in die Herausarbeitung des Grundgerüsts der HCI-Dialoge nach Kriterien der CA ein (vgl. Kapitel 4.1, *Rahmensequenzen*, *Komplettierungen*). Vor

dem Hintergrund dieses theoretischen Zugangs können mit Hilfe der annotierten Sprechakte pragmatische Muster, die typisch für die jeweiligen Anwendungen sind, herausgearbeitet werden. Die Annotation der Sprechakte ist außerdem besonders relevant im Rahmen der Diskussion um CT bei komplexeren Phänomenen, um Aussagen über funktionale Aspekte des UserInnen-Verhaltens stützen zu können. Dabei trägt sie z.B. zur quantitativen Analyse von komplexen Phänomenen wie sprachlicher Höflichkeit in der HCI bei.

3.3.6.2.5 Persistenzen

Um Alignment mit korpuslinguistischen Methoden nachweisen zu können, wurde für die vorliegende Studie *PerLexSy* entwickelt, eine Annotationskonvention für ***per***sistente ***lex***ikalische und ***sy***ntaktische Strukturen in Dialogen. Werden sprachliche Strukturen des Vorgänger-Turns im nächsten Turn oder weiteren Verlauf des Dialogs trotz SprecherInnen-Wechsel übernommen (Persistenz, Allo-Repetition), kann diese Beobachtung nach dem psycholinguistischen Modell von Pickering und Garrod (2004) als kognitives Alignment gedeutet werden, das auf der Ebene der Sprachverarbeitung von ihnen als *Lower-Level-Priming* gedeutet wird. Auf der Ebene der Performanz können keine Aussagen über Prozesse der Sprachverarbeitung im Gehirn getroffen werden. Mit empirischen Methoden lassen sich nur persistente Strukturen in Dialogen auffinden und klassifizieren nach ihren unterschiedlichen Persistenztypen und ihren unterschiedlichen Funktionen im Dialog. In *PerLexSy* kann bereits bei der Annotation zwischen lexikalischen und syntaktischen Persistenzen unterschieden werden (vgl. Pickering & Garrod 2004). Ferner wird zwischen Alpha- und Beta-Persistenz (vgl. Szmrecsanyi 2005) differenziert. Die Tags für Persistenzen stehen in eckigen Klammern. Große Initiale werden verwendet.

Um später Aussagen über Frequenzen und Funktionen der unterschiedlichen Persistenztypen treffen zu können, ist eine möglichst genaue Annotation unerlässlich. Persistenten Strukturen ist die Problematik inhärent, dass sie notwendig aus zwei oder mehr Teilen bestehen. Da beide Teile auch jeweils separat im Korpus aufgefunden werden können sollen, wird an jedem Teil der Persistenz eine Annotation vorgenommen.

(1) Twipsy: Genau! [La']
(2) User: genau [La]
(Twipsy 1508–1512)

Dieser Umstand ist nicht obligatorisch, denn man könnte auch vereinbaren, Persistenzen grundsätzlich am letzten Teil einer Paarsequenz zu taggen mit einem Tag für die gesamte Struktur.

(1) Twipsy: Genau!
(2) User: genau [La'-La]
(Twipsy 1508–1512)

Da bei der Auswertung von Persistenzen Paare gezählt werden und keine Einzelstrukturen, erleichtert die zweite Variante das automatische Auszählen. Die erste Variante hilft, die Strukturen im Korpus besser zurückverfolgen zu können. Sollen UserInnen- und Systemdaten separat ausgezählt werden, ist die jeweilige Annotation der Einzelstrukturen unerlässlich. Zusätzlich wurde innerhalb der Korpora eine farbige Markierung verwendet, um die zusammengehörigen Teile besser auffinden zu können.

Die Tags setzten sich nach folgendem Prinzip zusammen: „L" steht für *lexikalische Persistenz;* „S" steht für *syntaktische Persistenz*. Ein kleines „a" steht für *Alpha-Persistenz* und ein kleines „b" für *Beta-Persistenz*. Die vier Parameter können frei kombiniert werden. Zusätzlich wird jedes System-Turn-Tag mit einem Ableitungsstrich (') versehen zur besseren Unterscheidung. Denn es ist von großer Bedeutung, ob sich eine Userin persistent zum System verhält (Alignment?) oder sich das System zur Userin persistent verhält (Quasi-Persistenz aufgrund von Schlüsselworterkennung[102]). Entsprechend sind folgende Kombinationen möglich:

Tabelle 13: Tag-Set Persistenzkategorien (Allo-Repetition)

Kategorie	**Tag**	**Kommentar**	**Beispiel**
Lexikalische Alpha-Persistenz System-UserIn	**[La'-La]**	Die identische Wortform wird gespiegelt.	*System:* ***Schade!*** *UserIn:* ***schade***
Lexikalische Alpha-Persistenz UserIn-System	**[La-La']**	gleiche Kriterien wie oben, aber Interpretation als Quasi-Persistenz	*UserIn:* ***support*** *System: Hier auf Unserer* ***Support****-Seite sollten Sie alle gewünschten Informationen finden.*

102 Wenn das System eine persistente Struktur zum UserInnen-Input zeigt, so wird zunächst davon ausgegangen, dass diese ein Ergebnis des Schlüsselwort-Parsings sei und als Quasi-Pesistenz des Systems analysiert werden könne. Die Annotation wurde an dieser Stelle nicht bei jedem fraglichen Turn mit der konkreten Beschaffenheit des jeweiligen Dialog-Designs des fraglichen Systems abgeglichen. Für die vorliegende Studie standen keine umfassenden Informationen über die jeweiligen Makros zur Verfügung. Zur Weiterentwicklung innovativer Systeme wäre eine solche Evaluation aber hilfreich.

Kategorie	Tag	Kommentar	Beispiel
Syntaktische Alpha-Persistenz System-UserIn	**[Sa'-Sa]**	Die Struktur einer Phrase oder des ganzen Satzes wird übernommen.	*System:* ***Wer bist du?*** *<**ks'**>* *UserIn:* ***Ich bin Stefan.*** *<**ks**>*
Syntaktische Alpha-Persistenz UserIn-System	**[Sa-Sa']**	gleiche Kriterien wie oben, aber Interpretation als Quasi-Persistenz	*UserIn: Mich interessieren ihre Partner* ***in Österrreich****?* ***<pgr>*** *System: Bisher haben wir keine Partnerfirma* ***aus Österreich.*** ***<pgr'>***
Lexikalische Beta-Persistenz System-UserIn	**[Lb'-Lb]**	Variation in Form von • Komposita • Nominalisierungen • Personenwechsel bei Pronomina	*System: […]* ***Ausstellungsbereich*** *Kah Ih Robotik […]* *User: dann* ***Ausstellung*** *nicht machen.*
Lexikalische Beta-Persistenz UserIn-System	**[Lb-Lb']**	gleiche Kriterien wie oben, aber Interpretation als Quasi-Persistenz	*UserIn: Was gibt es heute* ***zu essen****.* *System:* ***Das Essen*** *in der Unimensa […]*
Syntaktische Beta-Persistenz System-UserIn	**[Sb'-Sb]**	• Strukturen mit gleichem Grundaufbau, aber unterschiedlichen Elementen • Identische Strukturen mit unterschiedlichen syntaktischen Funktionen	*System: Man* ***muss*** *doch nicht auf alles eine Antwort* ***haben****.* ***<vgr'>*** *(MV+VV)* *UserIn: Nee,* ***brauch*** *man nicht.* ***<vgr>*** *(MV+X)*
Syntaktische Beta-Persistenz UserIn-System	**[Sb-Sb']**	Gleiche Kriterien wie oben, aber Interpretation als Quasi-Persistenz	*UserIn: Sing ein Lied für mich!* ***<s>*** ***(VVImp+Akk+PGr)*** *System: Sing Du mir doch was vor.* ***<s>*** ***(VVImp+Nom+Dat+Akk)***

1. Lexikalisches und syntaktisches Alignment: Ein grundsätzliches Problem liegt darin, dass lexikalisches und syntaktisches Alignment häufig gemeinsam auftreten. Persistente Strukturen sind oft komplex, so dass bei einer Strukturübernahme mehrere Wortformen gespiegelt werden. Zusätzlich kann dabei auch die syntaktische Struktur übernommen werden.

(1) System: Ich sehe aus wie James Bond.
(2) User: Du siehst aus wie Stefan.
(Max-Korpus 1824–1836)

Eine solche Struktur kann unterschiedlich ausgewertet werden. Man könnte sie als eine einzige Konstruktion mit Platzhaltern betrachten, die in ihrer Gesamtheit übernommen wurde: „X sieht aus wie Y."

Entsprechend ließe sie sich nicht weiter in kleinere Bestandteile zerlegen und müsste mit einem einzigen Tag versehen werden. Dem Konzept von Alignment als Lower-Level-Priming und den Grundannahmen der Konstruktionsgrammatik käme eine solche Herangehensweise sehr nahe. Eine Konstruktion gewissermaßen als Chunk zu betrachten, der vom Menschen perzipiert und über Alignment Channels direkt reproduziert wird, entspräche der Modellvorstellung. Für die vorliegende Studie jedoch wurde eine Hilfskonstruktion entwickelt, um möglichst viele Aspekte einer Persistenz ermitteln zu können. So wurde die Persistenzrelation auf der Ebene der Gesamtkonstruktion in ihre einzelnen Bestandteile aufgeteilt und jeder Bestandteil wurde separat annotiert. Auf diese Art und Weise konnten auch alle untergeordneten Persistenzrelationen erhoben werden und nach den Parametern *lexikalisch* vs. *syntaktisch* und *Alpha* vs. *Beta* unterschieden werden.

Das Beispiel müsste dann so annotiert werden:

Tabelle 14: Beispiel Annotation

System	**User**	**Synt. Kategorie**	**Persistenzrelation**
ich	Du		[Lb'-Lb]
sehe	siehst		[La'-La]
aus	aus		
wie	wie		[La'-La]
James Bond	Stefan		
.	.	<s>	[Sa'-Sa]

Dass eine solche Klassifizierung auf der Mikroebene ausschließlich der detaillierten Erhebung von strukturellen Informationen dient, muss forthin auf Grundlage der vorliegenden Annotationskonvention bei jedem weiteren Analyseschritt mitgedacht werden.

Eine lexikalische Persistenz wird also zusätzlich getaggt, wenn sie Bestandteil einer komplett übernommenen syntaktischen Konstruktion ist. Für die Statistik erhalten wir so z. B. für die Persistenzrelation La'-La sowohl Werte von

ungebundenen Persistenzen als auch von syntaktisch gebundenen. Solche Zusatzinformationen müssen natürlich in die qualitative Auswertung miteinfließen.

2. **Auto-Repetition:** Auch Persistenzen zu eigenen Vorgänger-Turns (Auto-Repetition) können annotiert werden. Diese werden in der Psycholinguistik als Self-Alignment analysiert (vgl. Kapitel 2.3).
   Es gelten die gleichen Kriterien und Konventionen zur Annotation und die gleichen Tags werden verwendet. Es findet lediglich kein SprecherInnen-Wechsel statt. Entsprechend sind folgende Kombinationen möglich:

   UserIn: [La-La], [Lb-Lb], [Sa-Sa], [Sb-Sb]
   System: [La'-La'], [Lb'-Lb'], [Sa'-Sa'], [Sb'-Sb']

3. **Non-Persistenz:** Mit Bezug auf Szmrecsanyi (2005) wurde bei der Analyse eine zusätzliche Perspektive eingenommen: Es wurden nicht nur persistente Strukturen getaggt, sondern auch umgekehrt die Fälle, bei denen es zwar möglich gewesen wäre, eine Persistenz zu produzieren, diese aber im Korpus nicht auftaucht, sondern stattdessen ein Synonym oder eine Paraphrase. Selbstverständlich können auf diese Art und Weise ausschließlich Sequenzen in die Analyse eingehen, bei denen symmetrische Adjazenzpaare sehr wahrscheinlich sind. Die Entscheidung fiel hier theoriegeleitet auf ritualisierte Paarsequenzen im Gesprächsrahmen und Höflichkeitsmarker wie z.B. Duzen vs. Siezen (vgl. Kapitel 2.3). Wo also an diesen Loci keine Persistenzpaare produziert wurden, wurde die asymmetrische Form als Non-Alignment analysiert (vgl. Pickering & Garrod 2004, Fischer im Druck: 45). Die Untersuchung beschränkt sich hier auf das Verhalten der UserInnen bzw. der Bibliotheksmitglieder in Bezug auf lexikalische Alpha-Persistenz. Alles andere wäre zu spekulativ, da man eine konkrete erwartbare Form vorhersagen können muss, die wahrscheinlich ist, aber nicht produziert wurde (vgl. 2.3).

Tabelle 15: Non-Persistenz (keine Allo-Repetition)

Kategorie	**Tag**	**Kommentar**	**Beispiel**
Non-Persistenz System-UserIn	**[Na'-Na]**	kein persistentes Verhalten bei Grußfloskeln oder anderen ritualisierten Sequenzen	*System: Guten Tag.* *UserIn: Hi!*

Die unterschiedlichen Persistenz-Typen (Alpha-, Beta- und Non-Persistenz) stellen einen ersten Versuch dar, den **Grad der Ähnlichkeit** zwischen Prime und Persistenz zu erfassen. Dies ist allerdings schwer zu systematisieren, da Ähnlichkeit auf unterschiedlichen Ebenen betrachtet werden kann und wahrscheinlich als Kontinuum zwischen identischer Übereinstimmung und dem völligen Fehlen übereinstimmender Faktoren vorliegt. Der Operationalisierungsvorschlag in der Tabelle (s.u.) muss folglich als tentativ betrachtet werden.

Tabelle 16: Ähnlichkeitsgrade

Persistenz-Typ	**Ebenen der Übereinstimmung**	**Sprachliche Form**
Alpha-Persistenz	**+ Struktur** **+ Funktion**	Genau die gleiche Wortform oder syntaktische Struktur wird gespiegelt.
Beta-Persistenz	**+/- Struktur** **- Funktion**	Variation der Struktur oder gleich Struktur bei abweichender syntaktischer Funktion
Non-Persistenz	**- Struktur** **- Funktion**	keine Übernahme

Problematisch an diesem Versuch der Systematisierung ist die fehlende Ausdifferenzierung der strukturellen oder relationalen Variationsmöglichkeiten. Außerdem kann bei Turn-Wechseln ohne Persistenzen nur dann von Non-Persistenz im engeren Sinne gesprochen werden, wenn die persistente Struktur wahrscheinlich gewesen wäre (vgl. Szmrecsanyi 2005).

3.3.6.2.6 Überprüfung des Tag-Sets

Die Tag-Sets zu den unterschiedlichen Analyse-Ebenen wurden mit zwei Studierendengruppen im BA Germanistik (28 Personen, 12 Personen; 3.-8. Semester) in Einführungsseminaren zur HCI 2009–2010 getestet. Nach einer detailierten Einführung in die linguistischen Überlegungen und den Aufbau der Tag-Sets (3 x 90 min.) waren die Studierenden in der Lage, mit den Tag-Sets zu arbeiten und kamen überwiegend zu den gleichen Annotationen. Da es sich um Tags handelt, welche die strukturelle Seite von Sprache betreffen, sind bei geschulten Annotierenden weniger Abweichungen zu erwarten als bei semantischen Taggings. In der Studierendengruppe wurden allerdings besonders im Bereich der Syntax viele Fehler gemacht, die auf mangelnde Kenntnis des Grammatikmodells

zurückzuführen sind. Das Tag-Set führt also bei unterschiedlichen Annnotierenden zu gleichen Ergebnissen, sofern diese linguistisch gut vorgebildet sind. Die Analyse-Kategorien eignen sich nicht für einen crowd-basierten Ansatz mit Laien-Annotierenden (vgl. *Social-Taggings* und *Folksonomies*). Vielmehr wird der Ansatz der Experten-Annotation verfolgt. Die Annotationen der Persistenzen in den Untersuchungskorpora wurden zusätzlich von zwei Hilfskräften direkt gegengeprüft und im Nachgang ggf. modifiziert.

3.3.7 Auswertung

Die Korpora wurden auf die verschiedenen Untersuchungsbereiche abgestimmt in unterschiedlicher Form aufbereitet. Für die Analyse der Untersuchungsparameter aus dem Bereich der Konversationsanalyse, zu Kohärenz und Kohäsion sowie zu CT wurde eine Auszählung der fraglichen Phänomene (auf Ebene der Wortformen, der Syntax, der Sprechakte und der Dialogstruktur) pro Teilkorpus jeweils separat für UserIn und System vorgenommen. Die Ergebnisse werden als relative Häufigkeiten mit Bezug auf die Größe des jeweiligen Teilkorpus' angegeben, aus Gründen, die im folgenden Unterkapitel genauer ausgeführt werden sollen. Auf eine weitere statistische Aufbereitung wurde bei diesen Untersuchungsparametern verzichtet, da sich die Anzahl der Teilkorpora nur auf N=10 beläuft.

Im Rahmen der Teilstudie zu Persistenzen als Indikatoren für Alignment wurden die Untersuchungsparameter zusätzlich paarweise pro Dialog ausgezählt. Auf der Grundlage dieser größeren Grundgesamtheit (zwischen 20 und 51 Dialoge pro Teilkorpus) konnten weitere statistische Verfahren gerechnet werden. Zusätzlich wurde die Distanz (in Wortformen) zwischen den persistenten Strukturen erhoben und eine Distanz-Frequenz-Analyse durchgeführt.

3.3.7.1 Allgemeine Probleme bei der feinkörnigen Analyse von Felddaten

In den vorangegangenen Unterkapiteln wurden die Zusammensetzung der Korpora sowie die verschiedenen Untersuchungsparameter aus den in Kapitel 2 dargestellten theoretischen Ansätzen genau beschrieben. Dass solche Daten aus dem Feld statistisch nicht leicht zu verarbeiten sind, liegt auf der Hand. Für die weitere Aufbereitung stellten sich zwei Hauptprobleme:

- die unterschiedliche Länge der Dialoge
- die große Anzahl der zu untersuchenden Variablen

Im Folgenden soll skizziert werden, welche Entscheidungen auf Grund der schwierigen Ausgangsdaten für die statistische Aufbereitung getroffen wurden.

Zunächst musste entschieden werden, ob die Auswertung auf Basis einzelner Dialoge oder ganzer Teilkorpora erfolgen sollte.[103] Die vorliegenden Korpora sind das Ergebnis eines Extraktions- und Bereinigungsprozesses, der zu unterschiedlich großen Stichproben geführt hat (s. o.). Da die Felddaten per se nur eine geringe Anzahl an ganzen Dialogen enthielten, sollten diese auch in vollem Umfang genutzt werden. Die Dialoge differieren aber sowohl korpusimmanent als auch im Vergleich zwischen den Korpora stark in ihrer Länge. Hinzu kommt, dass jedes Korpus für die Auszählung der Daten zusätzlich in ein Korpus mit den Beiträgen der UserInnen und eines mit dem Output des jeweiligen Systems unterteilt wurde. Da die Systeme in der Regel längere Beiträge posten, unterscheiden sich die UserInnen-Korpora von den System-Korpora in Bezug auf die Dialoglängen (Anzahl der Wortformen pro Dialog) stark. Die Tabelle zeigt die deskriptive Statistik der Verteilung der Wortformen auf die Dialoge pro Teilkorpus.

Tabelle 17: Deskriptive Statistik Wortformen pro Dialog

		Twipsy UserIn	Twipsy System	Karlbot UserIn	Karlbot System	Max UserIn	Max System	Elbot UserIn	Elbot System	Bib Auskunft	Bib Mitglied
N	**Gültig**	27	27	48	48	39	39	20	20	51	51
	Fehlend	24	24	3	3	12	12	31	31	0	0
Mittelwert		49,13	328,80	12,33	141,80	33,72	191,08	104,46	548,44	56,09	42,32
Median		31,33	209,67	9,40	108,10	23,55	133,45	76,08	399,42	47,88	36,12
Standard-abweichung		69,11	462,52	8,83	101,58	29,97	169,86	105,12	551,90	45,17	34,07
Spannweite		341	2284	45	512	118	668	401	2107	203	153
Minimum		11	71	3	31	4	20	11	60	10	8
Maximum		352	2355	47	544	122	689	413	2166	213	161

103 Vorstellbar wäre auch eine Auswertung nach Turns, um einen höheren Wert für N zu erreichen. Die Turn-Längen in den Untersuchungskorpora variieren ihrerseits aber ebenfalls stark. Auf Grund ihrer großen Anzahl mit bis zu 900 Turns pro Korpus würden diese Abweichungen aber nicht mehr so stark ins Gewicht fallen. Eine Datenmaske anzulegen, in der man z. B. für jeden Turn einen Einzelwert einträgt, wäre bei dieser großen Anzahl von Turns sehr aufwändig. Dabei bliebe es fraglich, ob dieser „Schachzug" zu interpretierbaren Ergebnissen führen würde. Die verwendete Software (SPSS und r) kann derart große Datenmengen nicht mehr zuverlässig verarbeiten.

Vergleicht man die jeweiligen Teilkorpora in Bezug auf die Anzahl der Wortformen pro Dialog, wird deutlich, dass die Mittelwerte stark voneinander abweichen. Ein Blick auf die äußerst diversen Spannweiten verdeutlicht, dass die Korpora auch in sich heterogen sind in Bezug auf die Längen der Einzeldialoge. Die großen Standardabweichungen innerhalb und zwischen den Korpora erschweren die weitere statistische Aufbereitung.

Die große Anzahl an linguistischen Untersuchungsparametern, die in Kapitel 2 und 3 aus den theoretischen Grundlagen abgeleitet wurden, führt auf Ebene der Statistik zu außergewöhnlich vielen Variablen, die weiterverarbeitet werden müssten. Eine Auszählung all dieser Variablen pro Dialog ist aufwändig und kann zu einem Bodeneffekt (in sehr vielen Fällen der Wert von 0) führen, da nicht jeder Untersuchungsparameter auch in jedem Dialog vorkommt.

Für die übergeordnete Fragestellung sind drei Vergleichsrelationen relevant:

- UserInnen – System
- HCI – HHC
- Twipsy – Karlbot – Max – Elbot

Diese Vergleiche können auf der Basis von Teilkorpora angestellt werden. Würde ein Vergleich der einzelnen UserInnen angestrebt (z .B. Analyse der UserInnen-Typen), müssten die Dialoge separat behandelt werden. Da aber keine Meta-Daten zu den jeweiligen UserInnen zur Verfügung stehen, wäre eine solche Auswertung wenig aussagekräftig. Stattdessen wird die UserInnen-Sprache für das jeweilige gesamte Teilkorpus mit der Systemsprache für das parallele Teilkorpus verglichen. Ferner wird ein Vergleich zwischen den vier Systemen untereinander und mit dem Vergleichskorpus angestrebt. Es müssen dazu also lediglich die 10 Teilkorpora untereinander verglichen werden.

Theoretische und methodische Gründe sprechen in Bezug auf die meisten Untersuchungsparameter also gleichermaßen für die Auswertung der Daten pro Teilkorpus als relative Häufigkeiten mit Bezug zur Korpusgröße.

3.3.7.2 Zählung nach Teilkorpora

Alle Untersuchungsparameter wurden zunächst pro Teilkorpus ausgezählt (für UserIn und System separat). Eine automatische Zählung in Access ergab die in Tabelle 12 angegebenen Korpusdaten. Die Auszählung und Sortierung der einzelnen

Annotationskategorien erfolgte in Excel mittels der Pivot-Tabellen-Funktion.[104] Die Analysekategorien wurden zur Überprüfung der Hypothesen aus Kapitel 2 theoriegeleitet zu entsprechenden Gruppen zusammengefasst. Die unterschiedlichen Gruppen wurden auf die jeweils aussagekräftigste Einheit als Grundgesamtheit bezogen.

Tabelle 18: Analysekategorien und Bezugsgrößen

Analysekategorie	**Bezugsgröße**
Lemmata	Gesamtanzahl der Wortformen im Teilkorpus
POS	Gesamtanzahl der Wortformen im Teilkorpus
Syntax	Gesamtanzahl der TCUs im Teilkorpus
Dialogstruktur nach SWBD-DAMSL	Gesamtanzahl der Turns im Teilkorpus (Sprechakte) Gesamtanzahl der Dialoge im Teilkorpus (Grußsequenzen) Gesamtanzahl der TCUs im Teilkorpus (Dialogstruktur)
Persistenzen	Gesamtanzahl der Turns im Teilkorpus

3.3.7.3 Analyse der Untersuchungsparameter aus der Konversationsanalyse, zu Kohärenz und Kohäsion sowie zu Computer-Talk

Aus den o. g. Gründen wird in Bezug auf die meisten Untersuchungsparameter (CA, Kohärenz, CT) eine Auswertung nach Teilkorpora vorgenommen und nicht nach Dialogen. Mit vier untersuchten Systemen plus dem Vergleichskorpus beläuft sich also die Anzahl der Teilkorpora nach Aufteilung in je ein UserInnen- und ein System-Korpus auf N=10. Die Werte werden als relative Häufigkeiten zur jeweiligen Bezugsgröße angegeben (Menge der Wortformen, der Turns oder der Dialoge) und einander vergleichend gegenübergestellt. Relative Häufigkeiten mit sinnvollen Bezugsgrößen werden aus Gründen der Anschaulichkeit in Prozent angegeben.

104 Da sich die Untersuchung auf strukturelle Aspekte konzentriert, wurde von der Erstellung von Wortlisten in einem Konkordanzprogramm (MonoConc, Wordsmith-Tools, ConCGramm etc.) abgesehen.

Tabelle 19: Organisation der Teilkorpora

	Variable 1	**Variable 2**	**…**
Twipsy UserIn			…
Twipsy System			…
Karlbot UserIn			…
Karlbot System			…
Max UserIn			…
Max System			…
Elbot UserIn			…
Elbot System			…
Bib Auskunft			…
Bib Mitglied			…

Da die untersuchten Systeme grundverschieden sind (Chatbot mit Schlüsselworterkennung, Chatbot mit Dialogregeln, Syntaxparser, etc. und ECA), sind die Daten in den Systemkorpora voneinander zu unterscheiden. Die UserInnen-Daten stammen aus der Interaktion mit den unterschiedlichen Systemen und sind daher auch voneinander getrennt zu betrachten. Die Daten aus dem Vergleichskorpus müssen ebenfalls separat betrachtet werden, da es sich dabei um chat-basierte HHC in Abgrenzung zur HCI handelt.

Da für den Vergleich nur N=10 Teilkorpora zur Verfügung stehen, muss dann von jeglicher Inferenzstatistik abgesehen werden. Bei einem höheren Wert für N könnte man ebenfalls mittels einer einfaktoriellen Varianzanalyse (ANOVA) die unterschiedlichen Teilkorpora in Bezug auf mehrere Variablen gleichzeitig vergleichen, sofern diese parametrisch vorliegen. Aufgrund der geringen Datenmenge und der außergewöhnlich vielen Variablen, ist aber davon abzuraten. Entsprechend dürfen hier auch keine Korrelationen gerechnet werden, sondern es bleibt bei einer qualitativ vergleichenden Gegenüberstellung der Ergebnisse aus den verschiedenen Teilkorpora.

3.3.7.4 Teilstudie Alignment: Die Verteilung der Persistenzen

3.3.7.4.1 Zählung nach Dialogen

Eine Ausnahme bilden die operationalisierten Analysekategorien, die aus dem interaktiven Alignment-Modell abgeleitet wurden: lexikalische und syntaktische Persistenzen. Diese können in nahezu jedem Dialog nachgewiesen werden. Die Hypothesen aus Kapitel 2 lassen sich größtenteils mittels statistischer Verfahren

pro Dialog überprüfen. Um Effekte der dynamischen Anpassung zwischen UserIn und System zu untersuchen ist eine Analyse nach Dialogen unumgänglich.

Als vorbewusst ausgelöste Handlung müssten Persistenzen seitens der UserInnen weniger stark abhängig sein von Effekten, die von den unterschiedlichen Systemen hervorgerufen werden (anthropomorphes Design, kohärente Dialogführung), so dass eine rein quantitative Analyse hier sinnvoll ist.

3.3.7.4.2 Deskriptive Statistik

Für die Analyse des interaktiven Alignments wurde auf Grund der theoretischen Überlegungen und der großen Menge an Persistenzen trotz der o. g. statistischen Bedenken eine Analyse nach Dialogen vorgenommen (N=183 UserIn + System, bzw. N=366 bei geteilten UserInnen- und System-Korpora). Dabei wurde durchgängig mit relativen Werten in Abhängigkeit zur Dialoglänge gerechnet. Für die Verteilung der lexikalischen und syntaktischen Persistenzen bei UserIn und Sytem wurden Median, Minimum und Maximum errechnet (vgl. Kapitel 4.3). Dabei wurden Alpha- und Beta-Persistenzen zusammengefasst.

3.3.7.4.3 Inferenzstatistik

Der Kolmogorov-Smirnov-Test ergab, dass die Werte nicht normalverteilt vorliegen, also wurde für alle weiteren Verfahren mit Rangplätzen gerechnet. Zum Vergleich mehrerer Variablen (syntaktische und lexikalische Persistenz jeweils seitens der UserInnen oder des Systems in den unterschiedlichen Korpora) wurden Mann-Whitney-U-Tests zum Vergleich der zentralen Tendenz anhand der mittleren Ränge aus zwei unabhängigen Stichproben durchgeführt (als Alternative zum T-Test).[105] Dabei wird bei vergleichbaren zentralen Tendenzen davon ausgegangen, dass sich die Werte in einer gemeinsamen Reihe mit Rangplätzen gleichmäßig verteilen. So können die Mediane von zwei unterschiedlichen Stichproben verglichen werden.

Es wurde ein 1-faktorielles Design angesetzt; d.h. in jedem Test wurden die Verteilungen der Persistenzen korpus-immanent hinsichtlich eines Faktors verglichen. In separaten Rechnungen wurde zum einen die Unterscheidung zwischen UserIn und System (Faktor „*Interagierende*“) zugrunde gelegt und zum anderen die Unterscheidung zwischen Lexik und Syntax (Faktor „*Linguistische Beschreibungsebene*“).

105 Beim Mann-Whitney-U-Test (auch Mann-Whitney-Wilcoxon, MWW) handelt es sich um einen klassischen Hypothesentest zum Vergleich von zwei unabhängigen, non-parametrischen Stichproben.

1-faktorielles Design:

a) **Test 1**
Faktor „*Interagierende*“
mit zwei Stufen: „*UserIn*“ vs. „*System*“

Tabelle 20: Design Test 1

	Twipsy Lexik	Twipsy Syntax	Kalbot Lexik	Karlbot Syntax	Max Lexik	Max Syntax	Elbot Lexik	Elbot Syntax	Bib Lexik	Bib Syntax
UserIn										
System										

b) **Test 2**
Faktor „*Linguistische Beschreibungsebene*“
mit zwei Stufen: „*lexikalische Persistenz*“ vs. „*syntaktische Persistenz*“

Tabelle 21: Design Test 2

	Twipsy UserIn	Twipsy System	Karlbot UserIn	Karlbot System	Max UserIn	Max System	Elbot UserIn	Elbot System	Bib UserIn	Bib System
Lexik										
Syntax										

Pro Tabelle wurden zehn separate Tests gerechnet, da ein korpusimmanenter Vergleich der einzelnen Dialoge angestrebt wurde und nicht der Vergleich zwischen den Gesamtkorpora.

Für die Stichproben *UserIn* und *System* bzw. *Lexik* und *Syntax* werden gemeinsame Rangreihen angenommen. Mittels der Teststatistik U wird überprüft, ob die Rangplätze in der gemeinsamen Rangreihe gleich verteilt sind.

UserIn System UserIn System UserIn System
UserIn UserIn UserIn System System System

Die Signifikanz wird berechnet, indem der kleinere U-Wert mit dem kritischen Wert auf der theoretischen U-Verteilung verglichen wird. Von Post-Hoc-Verfahren (paarweiser Vergleich mit Bonferroni-Korrektur) wurde auf Grund der non-parametrischen Verteilung abgesehen. Da sich in beiden Tests der im Rahmen der Theorie interessante Effekt gerade in einer *gleich* verteilten Rangreihe ausdrückt (Beibehaltung der Nullhypothese), ist die bei non-parametrischen Stichproben ohnehin kritische Bestimmung der Effektgrößen weniger relevant.

3.3.7.4.4 Distanz-Frequenz-Analyse

Zusätzlich wurden zur Untersuchung von Distanz-Frequenz-Effekten die Abstände zwischen persistenten Strukturen pro Teilkorpus ausgezählt (Anzahl der Wortformen zwischen Prime und Persistenz) und eine Analyse der absoluten Werte vorgenommen. Da die Dialoglänge für diese Werte eine untergeordnete Rolle spielt, muss nicht mit relativen Werten gerechnet werden. Die Auswertung erfolgt in Form von Funktionsgraphen (x = Distanz (Recency) in Wortformen, y = Frequenz (Frequency)) in Tokens. Die Ergebnisse werden mit der Vergessensfunktion für persistente Strukturen in HHC-Dialogen (vgl. Szmrecsanyi 2005) verglichen.

3.3.7.5 Vor- und Nachteile des methodischen Ansatzes

a) **Ökologische Validität:** Die methodischen Probleme (N=10, non-parametrische Verteilung) und liegen in der Beschaffenheit der Daten in Form von Logfile-Korpora aus dem Feld begründet und können statistisch nicht gelöst werden. Die Stärke dieser Daten liegt dagegen in ihrer hohen ökologischen Validität. Außerdem sind die Werte aus den Korpora auf große Datenmengen zurückzuführen und daher viel aussagekräftiger als Werte, die von einzelnen Versuchspersonen stammen. Dies hat zwar statistisch keine Relevanz, fällt bei näherer Betrachtung der Daten aber ins Auge. Denn trotz der geringen Datenmenge bei N=10 bilden die Werte das jeweilige theoretische Modell teilweise sehr gut ab. In Kapitel 4 sollen dementsprechend trotz der statistischen Probleme auch für die Bereiche CA, Kohärenz und CT relative Häufigkeiten und Tendenzen für Koinzidenzen berichtet werden.
b) **Aussagekraft über diachrone Entwicklungen:** Obwohl die Daten in eine diachrone Abfolge gebracht werden können und einen Zeitraum von 2000 bis 2006 abdecken, handelt es sich bei der vorliegenden Untersuchung nicht um eine Longitudinalstudie. Es wurde nicht das UserInnen-Verhalten gegenüber ein und demselben System über einen Zeitraum von sechs Jahren untersucht und auch nicht eine feste Gruppe von UserInnen über die Jahre verfolgt, sondern die Daten aus den unterschiedlichen Jahrgängen stammen von sehr verschiedenen Systemen (vgl. Kapitel 1) und UserInnen-Gruppen. Es wird auch nicht ein Systemtypus, der über die Jahre weiterentwickelt wurde, anhand seiner unterschiedlichen Versionen evaluiert. Vielmehr differieren die untersuchten Systeme grundsätzlich in ihren Architekturen und Dialog-Designs. Im Vergleich zu den einfachen Chatbots Twipsy und

Karlbot stellen Max und Elbot zwar Weiterentwicklungen dar, es wurden jedoch unterschiedliche Aspekte von den jeweiligen Forschungsgruppen weiterentwickelt. Entwicklungstendenzen können hier zwar technologiehistorisch nachgezeichnet werden, man kann aber auf dieser Basis keine diachrone Linie in der Entwicklung der UserInnen-Sprache postulieren.