



UNIVERSITÄT
DES
SAARLANDES

semv·x
semantic technologies and voice solutions

Disambiguierungsstrategien in Dialogsystemen

Bachelorarbeit

Fachrichtung Computerlinguistik

vorgelegt von

Lena Enzweiler

Bachem, 21. November 2014

Eidesstattliche Erklärung

Ich versichere, die Bachelorarbeit selbstständig und lediglich unter Benutzung der angegebenen Quellen und Hilfsmittel verfasst zu haben.

Ich erkläre weiterhin, dass die vorliegende Arbeit noch nicht im Rahmen eines anderen Prüfungsverfahrens eingereicht wurde.

Bachem, 21. November 2014

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich während der Anfertigung dieser Bachelorarbeit fachlich und persönlich unterstützt haben.

Ich möchte mich zunächst bei Herrn Prof. Dietrich Klakow für die Überlassung des interessanten Themas bedanken.

Mein Dank gilt ganz besonders der Semvox GmbH und ihren Mitarbeitern, ohne die diese Arbeit nicht möglich gewesen wäre. Insbesondere danke ich Pia Kuznik, Dr.-Ing. Markus Löckelt und Jan Schehl für die Bereitstellung dieses interessanten Themas, die ständig freundliche Hilfsbereitschaft und für all die nützlichen Tipps zur Anfertigung dieser Arbeit. Pia Kuznik möchte ich außerdem ganz herzlich für die nette und engagierte Betreuung meiner Bachelorarbeit bedanken.

Ein ganz besonderer Dank gilt allen Personen, die sich mir als Versuchsperson und Korrekturleser zur Verfügung gestellt haben.

Ganz besonders möchte ich mich bei Christine Braun bedanken, die mir durch kritisches Hinterfragen und konstruktive Kritik immer wieder wertvolle Hinweise gab. Weiter möchte ich mich für die nützlichen Tipps zur Gestaltung meiner Arbeit bedanken.

Bei Tobias Aggintus möchte ich mich herzlich für die stetige Aufmunterung, alltägliche Unterstützung und Hilfe während der gesamten Studienzeit bedanken.

Mein ganz besonderer Dank gilt abschließend meiner Familie, insbesondere meiner Mutter, meiner Schwester, Vitter und Katja für die moralische und finanzielle

Unterstützung während meines gesamten Studiums.

Meine Bachelorarbeit entstand im Zeitraum vom September 2014 bis Dezember 2014 bei SemVox GmbH unter der Leitung von Prof. Dietrich Klakow und der Betreuung von Pia Kuznik.

Inhaltsverzeichnis

1. Einleitung	2
2. Related Work	5
3. Tools	6
4. Disambiguierungsstrategien	7
4.1. Disambiguierung	8
4.2. Disambiguierung in der Sprachverarbeitung	8
4.3. 1. Strategie: Aggregierte Auswahl ohne Pause	9
4.4. 2. Strategie: Aggregierte Auswahl mit Pause	9
4.5. 3. Strategie: Sequentielle Auswahl	11
5. Versuch 1	12
5.1. Wizard-of-Oz	12
5.2. Testszenario	13
5.3. Versuchsaufbau	15
5.4. Versuchsdesign	17
5.5. Control Panel	19
5.6. Versuchspersonen	20
5.7. Auswertung	21
5.7.1. Gemessene Zeiten	22
5.7.2. Fragebogen	25
5.7.3. Task Completion	37
5.7.4. Dialogverhalten	39
5.8. Resultat	40

6. Versuch 2	42
6.1. Testszenario	42
6.2. Versuchsaufbau	45
6.3. Versuchsdesign	46
6.4. Control Panel	46
6.5. Versuchspersonen	47
6.6. Auswertung	48
6.6.1. Gemessene Zeiten	48
6.6.2. Fragebogen	51
6.6.3. Task Completion	57
6.6.4. Dialogverhalten	58
6.7. Resultat	59
7. Ergebnisse	60
8. Diskussion	63
A. Anhang	i
Abbildungsverzeichnis	ii
Tabellenverzeichnis	iii

Abstract

Die vorliegende Arbeit beschäftigt sich mit der Frage, welche Disambiguierungsstrategien in Sprachdialogsystemen für Benutzer bei hoher kognitiver Belastung am geeignetsten sind. Man fokussiert sich dabei auf Sprachdialogsysteme, welche speziell für die Bedienung während der Autofahrt konzipiert werden. Um der Frage der besten Disambiguierungsstrategie nachzugehen, werden in einem Wizard-of-Oz-Experiment Fahrszenarien simuliert, bei denen die Versuchspersonen mit einem Dialogsystem sprachlich interagieren. Dabei werden ambige Eingaben des Benutzers simuliert, worauf das System mit Disambiguierungsstrategien in Form von Nachfragen reagiert. Diese verlangen eine entsprechende Benutzerreaktion. Anhand der Versuchsergebnisse wird analysiert, welche Strategien für den Benutzer am einfachsten zu benutzen und effektivsten waren. Insgesamt werden drei Disambiguierungsstrategien verfolgt. Aggregierte Auswahl ohne Pause, aggregierte Auswahl mit Pause, sowie die sequentielle Auswahl. Bei der aggregierten Auswahl ohne Pause werden alle möglichen Interpretationen einer ambigen Spracheingabe nacheinander in einer Sprachausgabe ausgegeben. Die aggregierte Auswahl mit Pause gibt alle möglichen Interpretationen durchnummeriert und durch Pausen getrennt in einer Sprachausgabe aus. In der sequentiellen Ausgabe wird jede Interpretation in einer separaten Sprachausgabe formuliert.

1. Einleitung

Dialogsysteme im automobilen Bereich müssen so gestaltet werden, dass sie den Fahrer so wenig wie möglich vom Fahren ablenken und ihm so gut wie möglich assistieren. Die Herausforderung für einen Entwickler von Sprachdialogsystemen (Dialogdesigner) besteht daher darin, Sprachäußerungen raffiniert zu gestalten. Dabei müssen dem Benutzer alle relevanten Informationen in verständlicher Weise geliefert werden. Auf die Sprachausgaben sollte möglichst einfach geantwortet und Anfragen und Wünsche effizient übermitteln werden können. Die Funktionsweise eines Dialogsystems hängt von mehreren Komponenten ab, welche anhand der in Abbildung 1 dargestellten Funktionsweise der Semvox GmbH eigenen ODP S3 Plattform ¹ kurz erläutert werden.

Die ODP S3 Plattform ermöglicht die Umsetzung komplexer Sprachdialoge, indem Spracheingaben eines Benutzers als semantische Objekte behandelt werden. Dabei wird eine Spracheingabe mit einer Backus-Naur-Form (BNF) Grammatik interpretiert, welche alle möglichen Spracheingaben des Benutzers listet und semantischen Objekten zuweist. BNF ist eine Metasprache, mit welcher bestimmt werden kann, ob eine Spracheingabe valide ist ([McCracken et al.]). Semantische Objekte werden von einem Backend-Server verarbeitet, woraufhin eine passende Sprachausgabe generiert wird.

¹ <http://www.semvox.de/de/technologie/odp-s3.html>

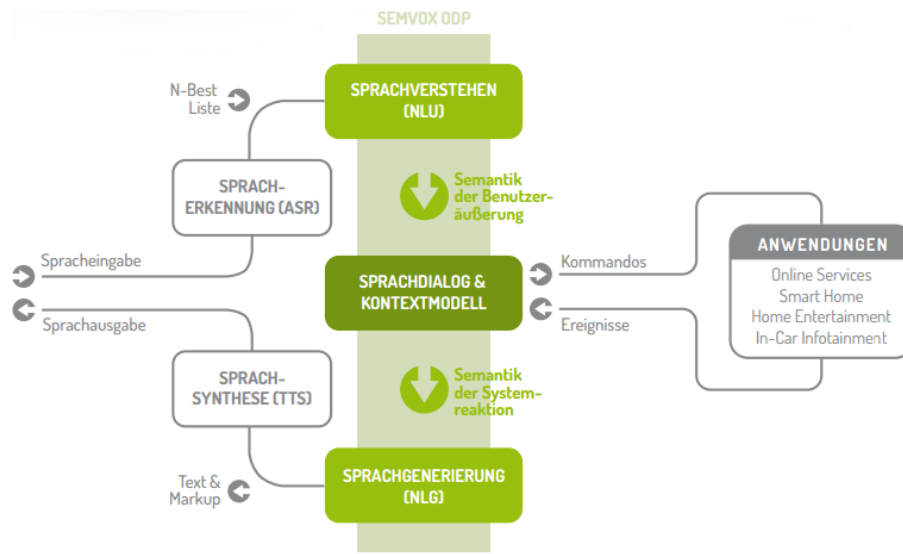


Abbildung 1: Funktionsweise der ODP S3 Plattform

In dieser Arbeit konzentriert man sich allein auf die Sprachgenerierung. Ein komplexer Dialog zwischen System und Benutzer führt häufig dazu, dass der Benutzer eine Eingabe macht, die das System nicht eindeutig zuordnen kann und mehrere Optionen für die Interpretation der vom Benutzer geäußerten Eingabe bestehen. Es muss an dieser Stelle vom System eine Rückfrage beim Benutzer erfolgen, sodass dieser seine vorherige Eingabe eindeutig übermitteln kann. Wenn der Benutzer zum Beispiel den Wunsch äußert, einen Kontakt aus dem im System gespeicherten Adressbuch anzurufen, es allerdings zwei Kontakte mit diesem Namen gibt, muss das System eine Rückfrage stellen, um zu ermitteln, welcher dieser beiden Kontakte gemeint ist. Der Dialogdesigner spricht in diesem Fall von Disambiguierung. Es wird in der vorliegenden Arbeit der Frage nachgegangen wie man konkret die Sprachausgabe einer solchen Disambiguierung innerhalb eines Dialogsystems, das speziell im Rahmen einer automobilen Anwendung konzipiert wird, gestaltet. Dabei werden drei verschiedene Strategien in einem Experiment

auf Effizienz und Beliebtheit unter den Versuchspersonen getestet. Diese Strategien werden im Kapitel 4 (Disambiguierungsstrategien) näher erläutert. Da für diesen Versuch kein Dialogsystem implementiert und lediglich ein Control Panel entwickelt wurde, welches die Sprachausgaben eines Systems simuliert, handelt es sich bei dem Versuch um ein Wizard-of-Oz Experiment. Dabei wird den Versuchspersonen mitgeteilt, dass sie mit einem echten Dialogsystem interagieren und sie wissen nicht, dass die Sprachausgaben vom Versuchsleiter gesteuert werden (siehe 5.1 Wizard-of-Oz). Der Versuch in Kapitel 5 (Versuch1) zeigt klar, dass bei einer Disambiguierung über wenige mögliche Slotfüller die Strategie **Aggregierte Auswahl ohne Pause** am beliebtesten unter den Versuchsperson ist. Es wurde sich daher für einen zweiten Versuch entschieden, welcher sich lediglich in der Länge der Disambiguierung unterscheidet. Dieser in Kapitel 6 (Versuch2) durchgeführte Versuch zeigt, dass bei einer Disambiguierung über mehrere mögliche Slotfüller die Strategie **Sequentielle Auswahl** die beliebteste Strategie ist. Neben der Beliebtheit unter den Versuchspersonen wurden bei beiden Versuchen weitere Faktoren, wie Dialogzeit, erfolgreiches Abschließen des Dialoges (Task Completion) oder Unterschiede des Dialogverhaltens zwischen hoher und geringer kognitiver Belastung erforscht. Ein zusammenfassendes Ergebnis beider Versuche findet sich in Kapitel 7 (Ergebnisse). Daran schließt sich eine Diskussion über die ermittelten Daten an, mit einem Ausblick für zukünftige Arbeiten.

2. Related Work

Dialogsysteme, Disambiguierungsstrategien in Interaktionen sowie kognitive Belastung von Videospielen in Verbindung mit Dialogsystemen sind Themen, die im Bereich der Computerlinguistik in diversen Arbeiten untersucht werden. Im Folgenden werden einige dieser Arbeiten vorgestellt.

Minker et al. untersuchten in einer Studie von 2012 eine weitere Disambiguierungsstrategie für Dialogsysteme. Übermittelt der Benutzer zum Beispiel bei der Navigation eine ambige Stadt als Zielort, fragt das System bei dieser Strategie nach dem ZIP-Code, um die Mehrdeutigkeit aufzulösen.

In den Studien von Tsiakoulis et al. von 2012 und Ang et al. von 2006 wurde die kognitive Belastung während Videospielen untersucht. Diese wurde weiter während einer Systeminteraktion von Tsiakoulis et al. von 2012 und Villing 2009 erforscht. Dabei wurde festgestellt, dass eine kognitive Belastung das Dialogverhalten ändert. Die Ergebnisse zeigen, dass während einer hohen kognitiven Belastung vom Benutzer längere Pausen zwischen zwei Sprachäußerungen eingelegt werden und die Anzahl der Sprachäußerungen geringer ist im Vergleich zur Anzahl während einer niedrigen kognitiven Belastung ([Villing, 2009]). [Tsiakoulis et al, 2012] kamen zu dem Ergebnis, dass das System während einer hohen kognitiven Belastung öfter durch sprachliche Äußerungen vom Benutzer unterbrochen wurde (Barge-in), als bei niedriger Belastung. In [Tsiakoulis et al, 2012] hat man weiter herausgefunden, dass kognitiv belastete Versuchspersonen Dialogabläufe mit einfachen Spracheingaben wie `ja` oder `nein` solchen Dialogabläufen bevorzugen, in denen das System den zu füllenden Slot als Antwort verlangt. Daher wird vermutet, dass die sequentielle Auswahl in dem hier vorgestellten Versuch bei Versuchspersonen mit hoher kognitiver Belastung am effizientesten ist. In der erwähnten Studie führen die Versuchspersonen ebenfalls parallel zur Dialoginteraktion ein Rennspiel. Man

hat dabei festgestellt, dass die Task Completion des aufgestellten Tasks bei der alleinigen Interaktion mit dem System höher war als bei der Interaktion parallel zum Rennspiel. Des Weiteren wurde herausgefunden, dass die Versuchspersonen einen Unterschied der kognitiven Belastung zwischen dem alleinigen Fahren, der alleinigen Interaktion und des Fahrens während der Interaktion festgestellt haben. Ähnliche Ergebnisse werden für die vorliegende Studie erwartet.

In Ang et al. von 2006 wurde erforscht, dass eine kognitive Belastung, die durch eine parallele Interaktion mit anderen Spielern in einem Computerspiel ausgelöst wird, die Leistung im Spiel verschlechtert. Die Kommunikation mit anderen Spielern im Computerspiel kann mit der Systeminteraktion aus dieser Studie verglichen werden. Deshalb wird für diese Studie eine schlechtere Rennspielleistung während des Testszenarios im Vergleich zur Rennspielleistung ohne Systeminteraktion erwartet. Dies könnte in zukünftigen Arbeiten überprüft werden.

In [Mishra et al., 2004] konnte festgestellt werden, dass Benutzer, die sich mehr auf einen anderen Task als auf die Systeminteraktion konzentrieren, eher unflüssige und abgehackte Sprachausgaben produzieren. In einer zukünftigen Arbeit könnte überprüft werden, ob solche Sprachäußerungen die angewendeten Disambiguierungsstrategien in einem echten System negativ beeinflussen. Des Weiteren kann diese Erkenntnis dazu genutzt werden, um die Stärke der Ablenkung durch das Rennspiel der einzelnen Versuchspersonen zu bewerten.

3. Tools

In diesem Kapitel sind die verwendeten Tools aufgelistet.

JavaFX² wurde zur Entwicklung eines Control Panels (siehe 5.5 Control Panel) verwendet, welches Sprachausgaben per Mausklick abspielt. Das Design wurde

² <http://docs.oracle.com/javase/8/javase-clienttechnologies.htm>

mit JavaFX Scene Builder³ entworfen. Die im Control Panel enthaltenen Sprachausgaben wurden online auf der Webseite <http://www.fromtexttospeech.com/> erstellt. Die Funktionen für das Abspielen und das Stoppen der Ausgaben wurden in Eclipse⁴ unter Installation des Addons E(fx)clipse⁵ implementiert.

Mit dem Rennspiel *Need for Speed: SHIFT*⁶ wurde in beiden Versuchen die Fahrsimulation realisiert.

Alle Fragebögen, die während des Versuchs abgefragt werden (siehe 5.7.2 Fragebogen) wurden mit Hilfe von Google Forms⁷ erstellt. Für die Auswertung der Antworten konnte die automatisch erstellte Zusammenfassung genutzt werden.

Um zu erforschen, ob die Ergebnisse des Versuchs statistisch signifikant sind, wurde das Programm InStat⁸ verwendet.

4. Disambiguierungsstrategien

Insgesamt werden in dieser Studie 3 Disambiguierungsstrategien auf Effizienz und Beliebtheit unter kognitiver Belastung getestet.

- Aggregierte Auswahl **ohne** Pause
- Aggregierte Auswahl **mit** Pause
- Sequentielle Auswahl

3 <http://www.oracle.com/technetwork/java/javase/downloads/javafxscenebuilder-info-2157684.html>

4 <https://www.eclipse.org/ide/>

5 <http://www.eclipse.org/efxclipse/index.html>

6 http://www.needforspeed.com/de_DE/shift

7 <http://www.google.com/forms/about>

8 <http://www.graphpad.com/scientific-software/instat>

In den folgenden Unterkapiteln wird zunächst kurz auf das Prinzip der Disambiguierung eingegangen. Anschließend werden die Funktionsweisen der einzelnen Strategien erläutert und mögliche Vor- und Nachteile, sowie Präferenzen der Versuchspersonen diskutiert.

4.1. Disambiguierung

Bei einer Disambiguierung werden verschiedene Begriffsbedeutungen voneinander abgegrenzt bzw. differenziert [Jurafsky et al.]. Dies gilt zum Beispiel für Nomen, welche den gleichen Begriff beschreiben aber ein anderes Konzept darstellen. Das Nomen *Bank* zum Beispiel kann sowohl ein Geldinstitut als auch eine Sitzmöglichkeit darstellen. Die Disambiguierung spielt bei der Sprachverarbeitung eine zentrale Rolle, da Spracheingaben nicht immer eindeutig formuliert werden und die dadurch entstehenden Mehrdeutigkeiten aufgelöst werden müssen [Jurafsky et al.].

4.2. Disambiguierung in der Sprachverarbeitung

Äußert ein Benutzer (im Folgendem auch User genannt) eines Dialogsystems eine ambige Spracheingabe, so muss das System diese disambiguieren. Diese Disambiguierung kann durch direkte Nachfrage der gewünschten Interpretation beim Benutzer erfolgen. Möchte der User zum Beispiel einen Kontakt aus seinem Adressbuch anrufen, dessen Vorname mehrfach vorkommt, so wird eine Disambiguierung notwendig sein, wenn der Benutzer bei seiner Spracheingabe lediglich den Vornamen angibt. Um den gewollten Kontakt vom User zu erfragen, kann das System eine der in dieser Arbeit behandelten Disambiguierungsstrategien verwenden.

4.3. 1. Strategie: Aggregierte Auswahl ohne Pause

Bei dieser Strategie werden alle möglichen Interpretationen der ambigen Spracheingabe in einer Sprachausgabe ohne Pause ausgegeben und auf eine Auswahl des Benutzers gewartet. In der folgenden Beispielinteraktion muss das System über den Nachnamen des von dem Benutzer adressierten Kontaktes disambiguieren. In der Sprachausgabe werden so alle möglichen Nachnamen (hier Meier und Müller) für den genannten Vornamen (hier Peter) zum Auswählen zur Verfügung gestellt. Der Benutzer kann während der Ausgabe mittels Barge-Ins antworten oder am Ende der Ausgabe mit dem gewünschten Nachnamen antworten.

Tabelle 1: Interaktionsbeispiel **Aggregierte Auswahl ohne Pause**

Akteur	Sprachausgabe
Benutzer	Rufe Peter an!
System	Meinst du Peter Müller oder Peter Meier?
Benutzer	Peter Müller.
System	Ok, ich werde Peter Müller jetzt anrufen.

Da diese Strategie einfach aufgebaut ist, sollte es für den Benutzer intuitiv klar sein, welche Antwort das System erwartet um die Interaktion weiter zu führen. Problematisch wird es bei einer hohen Anzahl an Disambiguierungsvorschlägen, da die Sprachausgabe entsprechend lang wird. Der Benutzer wird sich möglicherweise die komplette Sprachausgabe anhören, da die Möglichkeit zum Barge-In hier nicht auffällig ist.

4.4. 2. Strategie: Aggregierte Auswahl mit Pause

Diese Strategie funktioniert im Prinzip wie die 1. Strategie. Der Unterschied liegt darin, dass diese Strategie die einzelnen Vorschläge durchnummeriert präsentiert

und eine kurze Pause zwischen den Vorschlägen einlegt. Die Beispielinteraktionen zeigen die gleiche Situation wie zuvor. Allerdings antwortet der Benutzer im ersten Beispiel mit der Zahl, die der gewünschten Interpretation voran gestellt ist und im zweiten Beispiel mit Hilfe eines Barge-Ins.

Tabelle 2: Interaktionsbeispiel Aggregierte Auswahl mit Pause (Zahl)

Akteur	Sprachausgabe
Benutzer	Rufe Peter an!
System	Meinst du [Pause] 1. Peter Müller [Pause] oder 2. Peter Meier?
Benutzer	Den ersten.
System	Ok, ich werde Peter Müller jetzt anrufen.

Tabelle 3: Interaktionsbeispiel Aggregierte Auswahl mit Pause (Barge-In)

Akteur	Sprachausgabe
Benutzer	Rufe Peter an!
System	Meinst du [Pause] 1. Peter Müller [oder...]?
Benutzer	Ja.
System	Ok, ich werde Peter Müller jetzt anrufen.

Bei dieser Strategie ist die Möglichkeit zum Barge-In durch die Pausen auffälliger und der Benutzer muss dadurch nicht das Ende der kompletten Sprachausgabe abwarten, bevor er antwortet. Allerdings könnte die Sprachausgabe bei einer kleinen Anzahl an Disambiguierungsvorschlägen durch die Pausen und die Nummerierung unnötig lang auf den Benutzer wirken. Daher bevorzugt der Benutzer vermutlich

- die 1. Strategie bei einer kleinen Anzahl an Interpretation und entsprechend die 2. Strategie bei einer hohen Anzahl an Disambiguierungsvorschlägen.

4.5. 3. Strategie: Sequentielle Auswahl

Bei der Sequentiellen Auswahl steht jeder Disambiguierungsvorschlag in einer separaten Sprachausgabe und verlangt anschließend eine Bestätigung bzw. eine Ablehnung des angegebenen Vorschlages. Die ambige Spracheingabe wird dann mit der ersten Bestätigung des Benutzers aufgelöst. Das nachfolgende Beispiel entspricht der Interaktion aus Strategie 1 und 2 unter Berücksichtigung der Disambiguierungsstrategie 3.

Tabelle 4: Interaktionsbeispiel Sequentielle Auswahl

Akteur	Sprachausgabe
Benutzer	Rufe Peter an!
System	Meinst du Peter Meier?
Benutzer	Nein.
System	Meinst du Peter Müller?
Benutzer	Ja.
System	Ok, ich werde Peter Müller jetzt anrufen.

Diese Strategie ist wahrscheinlich besonders effizient, wenn der Benutzer einer hohen kognitiven Belastung ausgesetzt ist, da er das Tempo hier selbst bestimmen kann. Der Nachteil dieser Strategie liegt vermutlich darin, dass gerade bei vielen Interpretationsvorschlägen die Interaktion sehr lange dauert und der Benutzer jedes Mal eine Spracheingabe zur Fortsetzung des Dialoges eingeben muss.

5. Versuch 1

Um zu untersuchen, welche Disambiguierungsstrategie bei Versuchspersonen unter kognitiver Belastung am effizientesten ist, wird ein Wizard-of-Oz Experiment durchgeführt. Hierbei werden die Probanden ein Rennspiel fahren und parallel ein Testszenario durchführen, in welchem sie per Spracheingabe erfolgreich einen Anruf aufbauen sollen. Des Weiteren werden die Versuchspersonen dieses Testszenario ohne Rennspiel durchgehen, um mögliche Unterschiede der Ergebnisse zwischen kognitiv belastender und nicht kognitiv belastender Situation zu analysieren.

5.1. Wizard-of-Oz

In einem Wizard-of-Oz Experiment wird der Versuchsperson der Eindruck vermittelt, sie würde mit einem funktionierenden System interagieren. In Wirklichkeit wird die Existenz eines solchen Systems nur simuliert, in dem ein Versuchsleiter die Funktionen dieses mit einer entwickelten Software vortäuscht. ([Rogers et al.], [Jurafsky et al.]). Wizard-of-Oz Experimente werden generell dazu genutzt um eine Software vor der Implementierung zu testen [Jurafsky et al.].


In diesem Versuch wird ebenfalls ein Wizard-of-Oz Experiment durchgeführt, da man vor der Implementierung des Dialogsystems die effizienteste Disambiguierungsstrategie ermitteln möchte. Um ein fertig implementiertes System zu simulieren, wurde für diesen Versuch ein Control Panel entwickelt (siehe 5.5 Control Panel), mit dem Sprachausgaben eines fiktiven Dialogsystems vom Versuchsleiter simuliert werden können.

5.2. Testszenario

Während der Systeminteraktion sollen die Versuchspersonen erfolgreich einen Anruf aufbauen. Insgesamt sollen vier Personen angerufen werden, welche dem User über Personenprofile angezeigt werden. Darin sieht die Versuchsperson welche Slots zu füllen sind. Abbildung 2 zeigt das Personenprofil von Anke. Aus dem Profil geht hervor, dass Anke auf der geschäftlichen Festnetznummer angerufen werden soll.

Anke Schumacher



 Mainzerstr. 23, 66121, Saarbrücken

 A.Schumacher86@gmx.de

Abbildung 2: Personenprofil: Anke im 1. Versuch

Die Versuchspersonen werden am Anfang darauf hingewiesen, dass sie die Slots einzeln übergeben sollen. Nachdem der Benutzer per Sprachsteuerung spezifiziert hat, welche Person er anrufen möchte, fragt das System selbst die erforderlichen Slots ab. Diese Nachfrage wird für jeden der vier Anrufe in den unterschiedlichen Dialogstrategien realisiert. Pro Anruf gibt es insgesamt zwei zu füllende Slots, die mit derselben Disambiguierungsstrategie abgefragt werden. Die zu füllenden Slots sind in Tabelle 5 aufgelistet.

Tabelle 5: Beispiel Slotabfragen

Slot	erfragte Werte
Nummerntyp	Privat oder geschäftlich?
Telefontyp	Mobilnummer oder Festnetznummer
Nachname	Meier oder Müller
Stadt	München oder Ingolstadt

Welche Slots pro anzurufenden Kontakt abgefragt werden, zeigt Tabelle 6.

Tabelle 6: Slotabfrage pro Person

Anke	Peter	Fritz	Kim
Nummerntyp		Nummerntyp	Nummerntyp
Telefontyp	Telefontyp		Telefontyp
	Nachname		
		Stadt	

Damit man später die Dialogzeiten für jede Strategie vergleichen kann, soll bei jedem Anruf der Slot an zweiter Stelle der Disambiguierung gefüllt werden. Wann die zu füllenden Slots abgefragt werden, wissen die Versuchspersonen jedoch nicht.

5.3. Versuchsaufbau

Um eine möglichst realistische Fahrsimulation mit hoher kognitiver Belastung darzustellen, werden die Versuchspersonen ein Rennspiel spielen. Durch die Nutzung eines extra für Rennspiele konzipierten Lenkrads, inklusive Gas- und Bremspedal, entsteht ein realitätsgetreues Fahrgefühl. Bei dem Rennspiel handelt es sich um **Need for Speed: Shift**, welches im Einzelrennen - Modus mit jeweils fünf Gegnern gefahren wird. Abbildung 3 zeigt einen Ausschnitt des Rennspiels während des Versuchs.



Abbildung 3: Rennspiel während des Versuchs

Die Versuchspersonen sollen neben der Interaktion mit dem Dialogsystem auch eine möglichst hohe Platzierung erreichen. Dies soll die Konzentration und damit die kognitive Belastung während des Rennspiels steigern. Zu Beginn des Versuchs fahren die Probanden zunächst eine Testrunde. Das Ergebnis dieser Runde bietet eine Einschätzung darüber, wie schwer bzw. einfach das Rennspiel einer Testperson fiel. In den nächsten drei Runden werden die Versuchspersonen parallel zum Rennspiel das Testszenario durchgehen und dabei drei Personen anrufen.

Der Anruf gilt nur dann als erfolgreich, wenn alle Slots korrekt gefüllt werden. In der letzten Runde findet nur eine Systeminteraktion statt, ohne paralleles Rennspiel und die dadurch verursachte kognitive Belastung. Tabelle 7 zeigt eine Übersicht des Versuchsaufbaus.

Tabelle 7: Übersicht Versuchsablauf

Vorrunde	1. Runde	2. Runde	3. Runde	4. Runde
Rennspiel	Rennspiel	Rennspiel	Rennspiel	
	Anruf Anke	Anruf Peter	Anruf Fritz	Anruf Kim

Während des Versuchs werden die Versuchspersonen, das Rennspiel und die Dialoginteraktion aufgezeichnet. Dadurch wird sichergestellt, dass man alle Reaktionen einfangen und die Daten besser auswerten kann.

5.4. Versuchsdesign

Die Versuchspersonen fahren in den Runden 1-3 jeweils eine Strecke mit unterschiedlicher Disambiguierungsstrategie. Insgesamt werden diese auf drei verschiedenen Strecken verteilt, um einen Lerneffekt bei einer gleichbleibenden Strecke auszuschließen. Parallel werden die Zeiten gemessen, die eine Versuchsperson für die Absolvierung einer Strecke bei der Interaktion mit einer bestimmten Disambiguierungsstrategie benötigt (siehe Unterkapitel 5.7.1 Gemessene Zeiten). Da man diese Zeiten miteinander vergleichen möchte, müssen die Disambiguierungsstrategien geschickt auf die Strecken verteilt werden, da die Strecken unterschiedlich lang sind und daher keine aussagekräftigen Vergleiche untereinander bieten. Um diesen Konflikt zu lösen, werden die Versuchspersonen in drei Gruppen aufgeteilt, sodass jede Gruppe jede Strecke mit einer unterschiedlichen Disambiguierungsstrategie fährt. Schließlich kann man so für jede Strecke die Zeiten für unterschiedliche Strategien sammeln und vergleichen, mit welcher Strategie eine bestimmte Strecke am schnellsten gefahren wurde.

In der letzten Runde wird nur das Testszenario ohne Rennspiel durchgeführt. Hier-

für gibt es Gruppe 4, welche aus allen Versuchsteilnehmern besteht. Diese wird jedoch nochmal in drei Zwischengruppen aufgeteilt, sodass ein Drittel der Versuchspersonen in der vierten Runde das Testszenario in Strategie 1, ein Drittel in Strategie 2 und das letzte Drittel in Strategie 3 durchführt. Ein Überblick der Strecken- und Strategieverteilung pro Gruppe ist in Tabelle 8 aufgelistet.

Tabelle 8: Strecken- und Strategieverteilung

Aufteilung	Strategie 1	Strategie 2	Strategie 3
1. Gruppe	Strecke A	Strecke B	Strecke C
2. Gruppe	Strecke B	Strecke C	Strecke A
3. Gruppe	Strecke C	Strecke A	Strecke B
4. Gruppe	keine Strecke	keine Strecke	keine Strecke

Jede Gruppe fährt die Strecken in der gleichen Reihenfolge (erst Strecke A dann Strecke B und schließlich Strecke C). Dadurch wird gewährleistet, dass die Streckenzeiten durch keinen Lerneffekt bei einer unterschiedlichen Reihenfolge beeinflusst werden. Bei inkonsistenter Reihenfolge könnten die Resultate für die zuerst gefahrene Strecke aufgrund geringer Rennerfahrung schlechter ausfallen, als für die zuletzt gefahrene Strecke. Die anzurufenden Personen sind auf bestimmte Strecken festgelegt und in Tabelle 9 gelistet.

Tabelle 9: Anruf pro Strecke

Strecke	Anruf
Strecke A	Anke
Strecke B	Peter
Strecke C	Fritz
keine Strecke	Kim

5.5. Control Panel

Um ein laufendes System zu simulieren, wurde ein Control Panel entwickelt, welches verschiedene Sprachausgaben per Mausklick abspielen kann. Damit kann der Versuchsleiter die passenden Sprachausgaben auf entsprechende Benutzereingaben auslösen. Neben Ausgaben für die einzelnen Disambiguierungsstrategien sind weitere Sprachausgaben abgedeckt, welche oberflächlich zu jeder Eingabe des Benutzers eine Antwort bereitstellen und somit einen ungehinderten Ablauf des Dialogs gewährleisten. Zusätzlich dazu ist ein Stoppbutton enthalten, mit welchem per Klick alle aktiven Sprachausgaben abgebrochen werden können. Abbildung 4 zeigt das Control Panel. Für jede anzurufende Person gibt es ein extra Tab mit personenspezifischen Sprachausgaben. Die gemeinsamen Sprachausgaben wie **Cancel** und der Stoppbutton sind in jedem Personen-Tab extra enthalten, damit eine schnelle Reaktion des Versuchsleiters möglich ist. Das Commons-Tab enthält die Begrüßungsausgabe. Zur Orientierung ist nach jedem spezifischen Button die ausgelöste Sprachausgabe zu sehen.

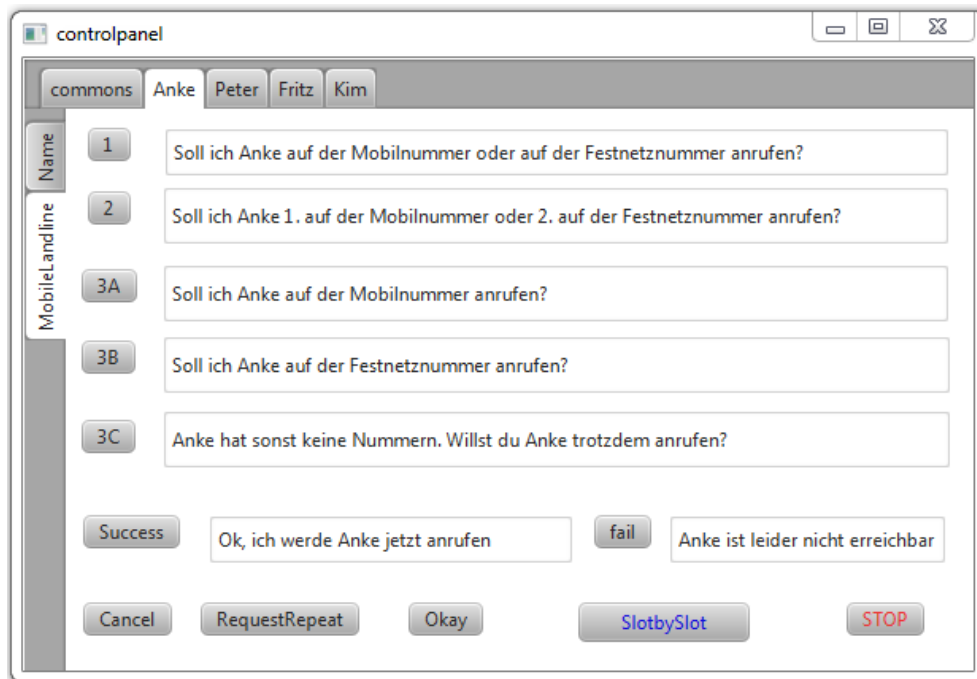


Abbildung 4: Controlpanel Versuch 1

5.6. Versuchspersonen

Es wurden 12 deutsche Muttersprachler in einer Alterspanne von 18-53 Jahren getestet. Zu Beginn wurden die Versuchspersonen gleichmäßig aber zufällig einer Gruppe zugeteilt, durch die bestimmt wird, welche Strategie auf welcher Strecke gefahren wird (siehe Tabelle 8: Strecken- und Strategieverteilung). Der Versuchsaufbau für die Versuchsperson sah folgendermaßen aus:

1. Testrunde fahren
2. Fragebogen über Person ausfüllen (siehe 5.7.2 Unterkapitel Person)
3. Strecke A fahren + Anke anrufen
4. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen (siehe 5.7.2 Unterkapitel Fragebogen)

5. Strecke B fahren + Peter anrufen
6. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen
7. Strecke C fahren + Fritz anrufen
8. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen
9. Kim anrufen
10. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen

Da den Versuchspersonen mitgeteilt wurde, dass sie mit einem echten System interagieren, sollten sie während des Dialogs deutlich in ein Tischmikrofon sprechen. Die Personenprofile konnten sie während des gesamten Dialoges über einen Laptop ansehen.

5.7. Auswertung

Um herauszufinden, welche Disambiguierungsstrategie am effizientesten ist, werden verschiedene Auswertungen vorgenommen. Zunächst werden die Zeiten gemessen, die die Versuchsperson zum einen für das Absolvieren der Strecke und zum anderen für das erfolgreiche Abschließen des Testszenarios benötigt (5.7.1 Gemessene Zeiten). Außerdem werden die Fragebögen ausgewertet, die von den Versuchspersonen nach jeder Runde ausgefüllt werden. Diese beziehen sich auf die subjektiv wahrgenommene kognitive Belastung und auf Merkmale der Disambiguierungsstrategien (5.7.2 Fragebogen). Des Weiteren wird die Task Completion ausgewertet, um zu erforschen, wie erfolgreich ein Dialog geführt wurde (5.7.3 Task Completion). Schließlich wird überprüft, wie die Versuchspersonen auf Rückfragen geantwortet haben und ob es dabei einen Unterschied zwischen hoch und niedrig belastenden Personen gibt (5.7.4 Dialogverhalten).

Um die Aussagekraft der Ergebnisse zu ermitteln, wird zusätzlich die statistische Signifikanz mit Hilfe des Tukey-Tests ermittelt.

5.7.1. Gemessene Zeiten

Rennzeiten

Um zu analysieren, ob das Rennverhalten durch eine Disambiguierungsstrategie negativ beeinflusst wird, werden die Rennzeiten pro Runde gemessen. Für jede Strecke wird die durchschnittliche Zeit gebildet, die die Versuchspersonen mit paralleler Systeminteraktion in einer bestimmten Strategie benötigten. Das Ergebnis ist in Tabelle 10 aufgelistet.

Tabelle 10: Durchschnittsrennzeiten jeder Strategie pro Strecke

Rennzeiten	Strategie 1	Strategie 2	Strategie 3
Strecke A	71,5 sek	93 sek	74,5 sek
Strecke B	68,75 sek	75,75 sek	91,5 sek
Strecke C	74,5 sek	58,38 sek	61,75 sek

Diesen Zeiten zufolge wurde jede Strecke mit einer bestimmten Strategie am besten gefahren. Strecke A und Strecke B wurde am schnellsten mit 71,5 Sekunden bzw. 68,75 Sekunden mit einer parallelen Dialoginteraktion in Strategie 1 gefahren. Mit Strategie 2 erfolgte die schnellste Rennzeit mit 58,38 Sekunden auf Strecke C. Diese Zeiten liegen jedoch nah beieinander und sind nicht aussagekräftig. Es gibt zudem keine Strategie, die durchweg auf allen Rennstrecken die besten Zeiten erzielen konnte. Dies könnte jedoch daran liegen, dass einzelne Werte durch schlechtere bzw. bessere Spieler in den Gruppen verfälscht wurden. Befindet sich zum Beispiel ein sehr schlechter Spieler in Gruppe 1 und ein sehr guter Spieler

in Gruppe 2, so könnte die Durchschnittszeit für Strategie 1 auf Strecke A, durch die lange Zeit des schlechten Spielers, verschlechtert werden. Im Gegensatz dazu könnte die Durchschnittszeit für Strategie 3 auf Strecke A durch die guten Resultate des guten Spielers aus Gruppe 3 verbessert werden. Um dieses Problem zu lösen, wird pro Strategie der Durchschnitt aller mit dieser Strategie gefahrenen Rennzeiten berechnet. Zeiten von extrem guten bzw. schlechten Spielern sollten die Durchschnittszeiten ganzer Strategien dann nicht mehr beeinflussen. Die daraus resultierenden Werte geben dann eine Aussage darüber, mit welcher Strategie die Rennen am besten bzw. am schlechtesten gefahren wurden. Die endgültige Rennzeitberechnung für die Analyse der effizientesten Disambiguierungsstrategie ist in Tabelle 11 dargestellt.

Tabelle 11: Durchschnittsrennzeiten pro Strategie

Rennzeiten	Strategie 1	Strategie 2	Strategie 3
Durchschnitt	71,58 sek	75,71 sek	75,92 sek

Die Unterschiede der Rennzeiten der einzelnen Strategien sind jedoch statistisch nicht signifikant. Daher kann hier nicht der Rückschluss gezogen werden, dass Strategie 1 wegen der insgesamt kürzesten Durchschnittsrennzeit von 71,58 Sekunden die Versuchspersonen am wenigstens ablenkt. Des Weiteren bleibt die Frage offen, ob die Rennzeiten überhaupt Ausschluss darüber geben können, welche Strategie sich am besten während einer tatsächlichen Autofahrt eignet. Dies könnte in zukünftigen Arbeiten durch einen umfangreicheren Versuch überprüft werden. Möglicherweise könnten besser Ergebnisse erzielt werden, wenn die Rennstrecken kürzer gewählt werden. Im aktuellen Versuch dauert eine Rennstrecke im Durchschnitt circa 74 Sekunden und ein Dialog 19 Sekunden. Das heißt, dass hier nur ca. ein Viertel der Rennstrecke mit paralleler Systeminteraktion gefahren wird.

Dialogzeiten

Neben den Zeiten für das Rennspiel werden auch die Dialogzeiten gemessen. Anhand dieser Zeiten kann man sehen, mit welcher Strategie der kürzeste Dialog möglich ist. Des Weiteren kann man durch einen Vergleich der Dialogzeiten, die Unterschiede in der Effizienz der Strategie mit und ohne kognitive Belastung durch das Autorennen sehen. Das könnte interessant sein, um die Unterschiede im Dialogverhalten zwischen einer kognitiv hoch belastenden Versuchsperson und einer weniger belastenden Person zu untersuchen. Eine längere Dialogzeit in einer gleichen Strategie ist möglicherweise auf eine längere Reaktionszeit zurückzuführen, weshalb bessere Zeiten in der vierten Runde, also ohne Rennspiel und damit ohne hohe kognitive Belastung, erwartet werden. Es werden alle Dialogzeiten aus den Runden mit Rennspiel gemessen. Aus diesen Zeiten lässt sich dann pro Strecke ein Durchschnittswert bilden. Neben diesen Werten wird zudem die gesamte durchschnittliche Dialogzeit pro Strategie gebildet. Diese Werte kann man mit den Durchschnittszeiten aus der Runde ohne Rennstrecke vergleichen. Es werden allerdings nur die Dialogzeiten bewertet, die einen korrekt durchgeführten Dialog abbilden, um die Durchschnittszeiten nicht zu verfälschen. Tabelle 12 zeigt diese Ergebnisse.

Tabelle 12: Durchschnittsdialogzeiten

Dialogzeiten	Strategie 1	Strategie 2	Strategie 3
Strecke A	15,34 sek	20,38 sek	20,28 sek
Strecke B	14,31 sek	20,05 sek	22,07 sek
Strecke C	15,97 sek	21,01 sek	20,35 sek
Strecke A - C	15,19 sek	20,52 sek	20,81 sek
ohne Strecke	14,9 sek	18,8 sek	17,59 sek

Die Unterschiede in den Dialogzeiten zwischen Strategie 1 und Strategie 2, sowie zwischen Strategie 1 und Strategie 3 sind statistisch signifikant. Die Unterschiede der Zeiten zwischen Strategie 2 und Strategie 3 sind nicht signifikant. Das Ergebnis zeigt, dass die Strategie 1 den kürzesten Dialog sowohl mit Rennspiel, als auch ohne Rennspiel ermöglicht. Die letzten beiden Zeilen der Tabelle zeigen, dass die Versuchspersonen einen kürzeren Dialog in jeder Strategie ohne Rennspiel ablegen. Die Ergebnisse aus 5.7.4 (Dialogverhalten) lassen ausschließen, dass die Unterschiede aufgrund unterschiedlichen Dialogverhaltens zustande kommen. Dies lässt vermuten, dass die Reaktionszeiten bei geringer Belastung schneller sind als bei hoher Belastung und die zeitlichen Unterschiede dadurch zustande kommen.

5.7.2. Fragebogen

Neben den Zeiten wird der Fragebogen jeder Runde ausgewertet. Dieser besteht im ersten Teil aus einem Ausschnitt des NASA-TLX Testes zur subjektiven Einschätzung der empfundenen kognitiven Belastung⁹. Im zweiten Teil werden Fragen über die zuletzt verwendete Dialogstrategie gestellt und es wird die Möglichkeit gegeben positives oder negatives Feedback über den Dialog der letzten Runde zu geben. Zu Beginn des Versuchs wird ein allgemeiner Fragebogen ausgefüllt, der Informationen zur Versuchsperson liefert.

Nasa-TLX

Abbildung 5 zeigt den NASA-TLX Fragebogen. Die Ergebnisse dieses Testes werden zum einen dafür genutzt um zu erforschen, bei welcher Strategie die Versuchspersonen eine höhere kognitive Belastung empfunden haben. Zum anderen kann man sehen, wie die Versuchspersonen ihre kognitive Belastung während einer Runde mit Rennspiel im Vergleich zur Runde ohne Rennspiel einschätzen.

⁹ http://www.keithv.com/software/nasatlx/nasatlx_german.html

Geistige Anforderung

Wie viel geistige Anforderung war bei der Informationsaufnahme und bei der Informationsverarbeitung erforderlich (z.B. Denken, Entscheiden, Rechnen, Erinnern, Hinsehen, Suchen ...)? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, erfordert sie hohe Genauigkeit oder ist sie fehlertolerant?

1 2 3 4 5 6

Gering ☐ ☐ ☐ ☐ ☐ ☐ Hoch

Körperliche Anforderung

Wie viel körperliche Aktivität war erforderlich (z.B. ziehen, drücken, drehen, steuern, aktivieren ...)? War die Aufgabe leicht oder schwer, einfach oder anstrengend, erholsam oder mühselig?

1 2 3 4 5 6

Gering ☐ ☐ ☐ ☐ ☐ ☐ Hoch

Zeitliche Anforderung

Wie viel Zeitdruck empfanden Sie hinsichtlich der Häufigkeit oder dem Takt mit dem die Aufgaben oder Aufgabenelemente auftraten? War die Aufgabe langsam und geruhsam oder schnell und hektisch?

1 2 3 4 5 6

Gering ☐ ☐ ☐ ☐ ☐ ☐ Hoch

Leistung

Wie erfolgreich haben Sie Ihrer Meinung nach die vom Versuchsleiter (oder Ihnen selbst) gesetzten Ziele erreicht? Wie zufrieden waren Sie mit Ihrer Leistung bei der Verfolgung dieser Ziele?

1 2 3 4 5 6

Gering ☐ ☐ ☐ ☐ ☐ ☐ Hoch

Anstrengung

Wie hart mussten Sie arbeiten, um Ihren Grad an Aufgabenerfüllung zu erreichen?

1 2 3 4 5 6

Gering ☐ ☐ ☐ ☐ ☐ ☐ Hoch

Frustration

Wie unsicher, entmutigt, irritiert, gestresst und verärgert (versus sicher, bestätigt, zufrieden, entspannt und zufrieden mit sich selbst) fühlten Sie sich während der Aufgabe?

1 2 3 4 5 6

Gering ☐ ☐ ☐ ☐ ☐ ☐ Hoch

Abbildung 5: Fragebogen: NASA-TLX

Die nachfolgende Tabelle zeigt die Ergebnisse des Fragebogens. Dabei werden jeweils die durchschnittlichen Antworten für alle Runden (1-4), der Runden mit Rennspiel (1-3) und der Runde ohne Rennspiel (4) aufgelistet und für jede Strategie einzeln gewertet.

Antwortenintervall	Strategien	Ergebnisse bestimmter Runden		
		1-4	1-3	4

Geistige Anforderung

1: gering 6: hoch	Strategie 1	1,88	2,08	1,25
	Strategie 2	2,06	2,42	1
	Strategie 3	2,63	2,83	2

Körperliche Anforderung

1: gering 6: hoch	Strategie 1	2	2,17	1,5
	Strategie 2	1,44	2,25	1
	Strategie 3	2	2,25	1,25

Zeitliche Anforderung

1: gering 6: hoch	Strategie 1	1,87	2,09	1,25
	Strategie 2	1,75	2	1
	Strategie 3	2,31	2,67	1,25

Leistung

1: gering 6: hoch	Strategie 1	4,25	4,5	3,5
	Strategie 2	4,75	4,33	6
	Strategie 3	4,25	4	5

Anstrengung

1: gering 6: hoch	Strategie 1	2	2,25	1,25
	Strategie 2	2,13	2,55	1
	Strategie 3	2,63	2,92	1,75

Frustration

1: gering 6: hoch	Strategie 1	1,69	1,83	1,25
	Strategie 2	1,81	2,08	1
	Strategie 3	2	2,25	1,25

Die Unterschiede der Antworten einzelner Strategien sind nicht signifikant, sodass nicht eindeutig gesagt werden kann, welche Strategie die Versuchsperson am meisten bzw. am wenigsten belastet. Betrachtet man jedoch die Ergebnisse aller Fragen fällt auf, dass die Antworten im Bezug auf geistige Anforderung, die Anstrengung und die Frustration zeigen, dass die erste Strategie als am wenigsten belastend und die dritte Strategie als am belastendsten empfunden wurde. Dies deckt sich auch mit den Ergebnissen aus dem zweiten Teil des Fragebogens, in welchem die erste Strategie als Favorit und die dritte Strategie als unbeliebteste Strategie gewertet wurde.

Vergleicht man die Werte für Runde 1-3 mit den Werten von Runde 4 fällt auf, dass die Dialoge parallel zum Rennspiel durchweg, mit Ausnahme der Frage nach der Leistung, als belastender gewertet wurden als die Dialoge ohne Rennspiel. Dies zeigt, dass die Versuchspersonen einen Unterschied in der kognitiven Belastung gespürt haben.

Strategien

Der zweite Teil des Fragebogens ist in Abbildung 6 zu sehen. Mit diesem Fragebogen werden die einzelnen Strategien anhand verschiedener Kategorien bewertet.

Dialogverhalten

Wie zufrieden waren sie mit der Systeminteraktion

Der Dialog lenkte mich stark vom Rennspiel ab

Fiel es Ihnen schwer, das Rennspiel parallel zur Systeminteraktion zu spielen und so eine gute Leistung zu absolvieren?

1 2 3 4 5 6

lenkte mich kaum ab ☐ ☐ ☐ ☐ ☐ ☐ lenkte mich stark ab

Die Systemnachfragen erleichterte es mir, den Anruf korrekt aufzubauen

Hat das System dir dabei geholfen, die richtigen Personendaten einzugeben und somit eine Person korrekt mit den vorgegebenen Angaben anzurufen?

1 2 3 4 5 6

erleichterte die Eingaben ☐ ☐ ☐ ☐ ☐ ☐ erschwerte die Eingaben

Wussten Sie, zu welchem Zeitpunkt das System Spracheingaben erwartete?

Haben Sie gemerkt, wann das System auf eine Spracheingaben von Ihnen wartet um den Dialog fortzuführen?

1 2 3 4 5

habe die Stellen immer erkannt ☐ ☐ ☐ ☐ ☐ habe die Stellen nicht immer erkannt

Wie gefiel Ihnen der Dialog insgesamt?

1 2 3 4 5 6

Sehr gut ☐ ☐ ☐ ☐ ☐ ☐ Weniger gut

Gab es etwas was Ihnen an dem Dialog sehr gut gefiel?

Gab es etwas was Ihnen an dem Dialog nicht gefiel?

Abbildung 6: Fragebogen: Dialogstrategien

Die nachfolgende Tabelle zeigt die Ergebnisse des Fragebogens. Dabei werden jeweils die durchschnittlichen Antworten für alle Runden (1-4), der Runden mit Rennspiel (1-3) und der Runde ohne Rennspiel (4) aufgelistet und für jede Strategie einzeln gewertet.

Da die Interaktion mit dem Dialogsystem in Runde 4 ohne Rennspiel erfolgt, wird die Frage "Der Dialog lenkte mich vom Rennspiel ab" für diese Runde nicht beantwortet. Entsprechend wird die Frage "Fiel es Ihnen einfacher, den Dialog ohne Rennspiel zu führen?" nur für die 4. Runde beantwortet. Die Frage "Welcher Anruf gefiel Ihnen insgesamt am besten?" wird zum Schluss beantwortet.

Antwortenintervall	Strategien	Ergebnisse bestimmter Runden		
		1-4	1-3	4

Der Dialog lenkte mich vom Rennspiel ab

1: kaum 6: stark	Strategie 1	-	2,25	-
	Strategie 2	-	2,58	-
	Strategie 3	-	2,58	-

Die Nachfragen erleichterten es mir, den Anruf korrekt aufzubauen

1: erleichterte es 6: erschwerte es	Strategie 1	1,63	1,83	1,25
	Strategie 2	1,69	1,92	1
	Strategie 3	2,13	2,25	1,75

Wussten Sie, wann das System Spracheingaben erwartete?

1: immer 6: nicht immer	Strategie 1	1,19	1,17	1,25
	Strategie 2	1,38	1,33	1,5
	Strategie 3	1,5	1,67	1

Wie gefiel Ihnen der Dialog insgesamt?

1: sehr gut 6: weniger gut	Strategie 1	1,94	2,08	1,5
	Strategie 2	2,50	2,67	2
	Strategie 3	2,57	2,75	2

Fiel es Ihnen einfacher, den Dialog ohne Rennspiel zu führen?

1: viel einfacher 6: nicht einfacher	Strategie 1	-	-	2
	Strategie 2	-	-	3,75
	Strategie 3	-	-	2,25

Welcher Anruf gefiel Ihnen insgesamt am besten?

Anruf bzw. Strategie auswählbar	Strategie 1	75%	-	-
	Strategie 2	16,6%	-	-
	Strategie 3	8,3%	-	-

Die ersten vier Antworten dieses Fragebogens zeigen, dass die erste Strategie am Positivsten und die dritte Strategie am Negativsten gewertet wurde. Dies stimmt mit dem Ergebnis der letzten Frage überein, welche konkret nach der beliebtesten Strategie fragt.

Die vorletzte Frage zeigt, dass es dem Durchschnitt der Versuchspersonen einfacher fiel, den Dialog ohne Rennspiel zu führen. Dies bestätigt das Ergebnis aus dem NASA-TLX Fragebogen, welches besagt, dass die Versuchspersonen einen Unterschied in der kognitiven Belastung zwischen Dialog mit Rennspiel und ohne Rennspiel gemerkt haben.

Person

In Abbildung 7 sind die Fragen dieses Fragebogens abgebildet. Dieser Fragebogen dient dazu, um Informationen über die Versuchsperson zu erhalten. Die Fragen nach der Rennspiel- und Dialogerfahrung können für die spätere Auswertung der Zeiten interessant sein und eine mögliche Erklärung für stark abweichende Rennspiel- und Dialogzeiten liefern.

Jeder Person wird eine ID zugeteilt. Diese wird zu Beginn jedes Fragebogens eingetragen, damit jeder Versuchsperson alle abgegebenen Antworten zugeordnet werden können.

Fragen zur Person

Wie ist ihre ID?

Wie alt sind Sie?

Haben Sie Erfahrung mit Dialogsystemen?

1 2 3 4 5 6

gar keine Erfahrung ☐ ☐ ☐ ☐ ☐ ☐ viel Erfahrung

Spielen Sie oft Rennspiele?

1 2 3 4 5 6

sehr oft ☐ ☐ ☐ ☐ ☐ ☐ nie

Wie technikaffin sind Sie?

1 2 3 4 5 6

sehr technikaffin ☐ ☐ ☐ ☐ ☐ ☐ gar nicht technikaffin

Wie schwer fiel Ihnen die Einführungsrunde?

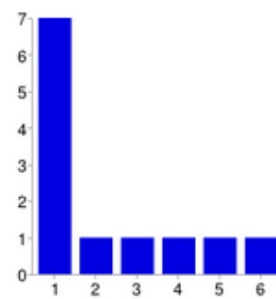
1 2 3 4 5 6

sehr schwer ☐ ☐ ☐ ☐ ☐ ☐ sehr einfach

Abbildung 7: Fragebogen: Person

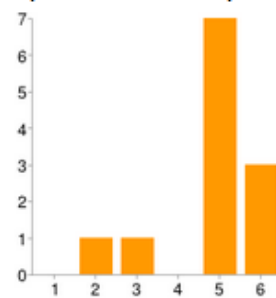
58% der Probanden sind in einer Altersgruppe von 18-29, 17 % in einer Altersgruppe von 30-41 und 25% in einer Altersgruppe von 42-53. Alle Versuchspersonen sind deutsche Muttersprachler. Eine Zusammenfassung der restlichen Antworten steht in Abbildung 8.

Haben Sie Erfahrung mit Dialogsystemen?



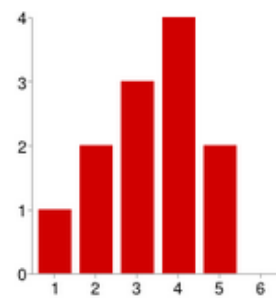
1	7	58 %
2	1	8 %
3	1	8 %
4	1	8 %
5	1	8 %
6	1	8 %

Spielen Sie oft Rennspiele?



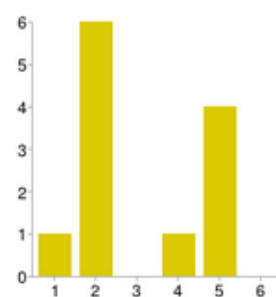
1	0	0 %
2	1	8 %
3	1	8 %
4	0	0 %
5	7	58 %
6	3	25 %

Wie technikaffin sind Sie?



1	1	8 %
2	2	17 %
3	3	25 %
4	4	33 %
5	2	17 %
6	0	0 %

Wie schwer fiel Ihnen die Einführungsrunde?



1	1	8 %
2	6	50 %
3	0	0 %
4	1	8 %
5	4	33 %
6	0	0 %

Abbildung 8: Fragebogen: Person

Das Ergebnis zeigt, dass 75% der Testpersonen keine bzw. wenig Erfahrung mit Dialogsystemen haben. Aus den Antworten lässt sich weiter schließen, dass 83% der Befragten selten Rennspiele spielen und die Einführungsrunde 58% schwer fiel. Daraus wird klar, dass die Mehrheit der Versuchspersonen unerfahren im Umgang mit Dialogsystemen und Rennspielen ist. In einer zukünftigen Arbeit könnte untersucht werden, ob das Ergebnisse des Versuchs abhängig von der Erfahrung der Versuchspersonen mit Dialogsystemen und Rennspielen ist. Dazu könnten die Testpersonen je nach Erfahrung in zwei Gruppen aufgeteilt und die Resultate miteinander verglichen werden.

5.7.3. Task Completion

Für jede Strategie wird die Task Completion ausgewertet, welche besagt, mit welchem Erfolg das Testszenario ausgeführt wurde. Sie wird bemessen, in dem man für jeden richtig gefüllten Slot (siehe Tabelle 6) einen Punkt verteilt. Folgenden Punktzahlen sind also für jede Strategie möglich:

- 0 Punkte, wenn kein Slot richtig gefüllt wird
- 1 Punkt, wenn ein Slot richtig gefüllt wird
- 2 Punkte, wenn alle Slots richtig gefüllt werden

Zur Auswertung wird pro Strategie eine Durchschnittspunktzahl berechnet. Die Durchschnittspunktzahlen für alle Runden (1-4), der Runden mit Rennspiel (1-3) und der Runde ohne Rennspiel (4) finden sich in Tabelle 15.

Tabelle 15: Durchschnittliche Task Completion pro Strategie

Strategien	insgesamt	Runde 1-3	Runde 4
1. Strategie	1,75	1,92	1,5
2. Strategie	1,94	1,92	2
3. Strategie	1,63	1,5	2
Insgesamt		1,78	1,83

Im Durchschnitt wurde in Runde 4 eine höhere Task Completion erreicht. Das besagt, dass die Dialoge der vierten Runde am erfolgreichsten durchgeführt wurden. Der Unterschied ist jedoch sehr gering, sodass man für ein eindeutiges Resultat mehr Ergebnisse zur Auswertung benötigt. Die Durchschnittspunktzahlen zeigen weiter, dass die zweite Strategie insgesamt am erfolgreichsten abgeschlossen wurde. Die erfolgloseste Strategie ist nach diesem Ergebnis die dritte Strategie. Dies passt zum Ergebnis, dass diese Dialogstrategie im Fragebogen über die Strategien (vgl. 5.7.2 Unterkapitel Strategien) am schlechtesten gewertet wurde. Strategie 1, welche als beliebteste Strategie der Runde 1-3 ausgewählt wurde, liefert für diese Runden die gleiche Task Completion wie Strategie 2. Hier fehlen weitere Ergebnisse um eine konkrete Verbindung zwischen beliebteste Strategie und Strategie mit höchster Task Completion herzustellen. Dabei muss auch der Fall betrachtet werden, dass die Versuchsperson ihre Fehler möglicherweise gar nicht bemerken. Diese Verbindung könnte in einem umfangreichen Experiment in späteren Arbeiten überprüft werden. Grundsätzlich kann man an diesem Ergebnis sehen, dass insgesamt der Anruf am häufigsten korrekt mit Strategie 2 und am seltensten korrekt mit Strategie 3 ausgeführt werden konnte.

Es ist jedoch fraglich, ob die hier entstandenen Fehler auch in einem realen Dialog aufkommen, bei dem die Versuchsperson die Anrufattribute selbst bestimmt. Des-

halb gibt diese Auswertung nur ein Indiz darauf, welche Strategie möglicherweise am kompliziertesten ist.

5.7.4. Dialogverhalten

Für jede Strategie werden die Antworten der Versuchspersonen gesammelt um festzustellen, ob es Unterschiede im Dialogverhalten unter hoher und niedriger kognitiver Belastung gibt. Alle gegebenen Antwortmöglichkeiten werden kurz erläutert und sind in Tabelle 16 mit einem Beispiel aufgelistet.

Slots: der zu füllende Slot wird als Antwort übergeben.

Position: es wird mit der Position des gewünschten Slotfüllers geantwortet.

ja/nein: der vorgeschlagenen Slotfüller wird angenommen bzw. abgelehnt.

Tabelle 16: Antwortmöglichkeiten

Antwort- möglichkeiten	Beispiel Slotabfrage	Beispiel Antwort
Slots	Willst du Anke privat oder geschäftlich anrufen?	geschäftlich
Position	Willst du Anke 1. privat oder 2. geschäftlich anrufen?	zweitens
ja/nein	Willst du Anke privat anrufen?	nein

Die Häufigkeiten dieser Antworten pro Strategie aus Runde 1-3 sind in nachfolgender Tabelle aufgelistet.

Tabelle 17: Antwortenverteilung pro Strategie

Antwort- möglichkeiten	Strategie 1	Strategie 2	Strategie 3
Slots	100%	70,8%	16,7%
Position	0%	29,2%	0%
ja/nein	0%	0%	83,3 %

Um zu erforschen, ob sich das Dialogverhalten in Runde 4 ändert, wurden pro Person die Antworten aus der Strategie der 4. Runde mit der entsprechenden Strategie aus den Runden davor verglichen. Es hat sich herausgestellt, dass die Verteilung der Antwortmöglichkeiten bei hoher Belastung die gleiche ist wie bei niedriger Belastung. Dadurch ist kein Unterschied im Dialogverhalten bei unterschiedlicher Belastung erkennbar.

5.8. Resultat

Aus den Resultaten aus Kapitel 5.7 (Auswertung) wird die effizienteste Strategie ermittelt.

Die Ergebnisse aus den Rennzeiten zeigen, dass die Rennstrecken mit Strategie 1 am Schnellsten befahren wurden. Dieses Resultat ist jedoch nicht verlässlich, da die Werte statistisch nicht signifikant sind. Die erzielten Dialogzeiten zeigen deutlich, dass Strategie 1 sowohl in den Runden mit als auch ohne Rennspiel den kürzesten Dialog ermöglicht. Aus den Antworten des NASA-TLX Teil des Fragebogens wird

deutlich, dass Strategie 1 von den Benutzern als am wenigsten belastend gewertet wurde. In allen Fragen des zweiten Teils des Fragebogens wurde ebenfalls Strategie 1 am besten bewertet und durch die letzte Frage deutlich als beliebteste Strategie gewertet. Laut Task Completion ist Strategie 2 insgesamt am erfolgreichsten, Strategie 1 und 2 in den Runden mit Rennstrecke jedoch gleich gut. Durch diese Erkenntnisse kommt man zu dem Entschluss, dass Strategie 1 die beliebteste und effizienteste Strategie ist.

Da dieses Resultat bereits nach wenigen Versuchspersonen zu erwarten war und die Frage aufkam, ob die erste Strategie auch bei einer längeren Disambiguierung am geeignetsten ist, hat man den Versuch bereits nach 12 Versuchspersonen abgebrochen und einen zweiten Versuch gestartet. Der zweite Versuch ist identisch zum Aufbau des ersten Versuchs, unterscheidet sich jedoch in der Anzahl der in der Disambiguierung vorgeschlagenen Slotfüller.

6. Versuch 2

Da die Ergebnisse des ersten Versuches sehr einheitlich gezeigt haben, dass bei einer Disambiguierung über zwei Füllslots (zum Beispiel: Peter Müller oder Peter Meier) die erste Strategie am effizientesten und beliebtesten ist, hat man sich zusätzlich für einen weiteren Versuch entschieden. In diesem Versuch werden pro Disambiguierung mehr als zwei mögliche Füllslots vorgeschlagen (zum Beispiel: Peter Müller, Peter Meier, Peter Lauer, Peter Fischer, Peter Schneider oder Peter Schmidt). Damit will man herausfinden, ob die erste Strategie auch bei längerer Disambiguierung bevorzugt wird.

6.1. Testszenario

Dieses Testszenario ist analog zu dem im ersten Versuch. Die Versuchspersonen rufen in den ersten drei Runden jeweils Anke, Peter und Fritz an. Dabei fahren sie parallel zur Interaktion mit dem System ein Rennspiel. In der vierten Runde wird nur Kim angerufen. Dies geschieht ohne Rennspiel. Der Versuch unterscheidet sich jedoch in den zu füllenden Slots, welche in Tabelle 18 aufgelistet sind. Die Anzahl der vorgeschlagenen Füllern für den jeweiligen Slot ist in Klammern angegeben.

Tabelle 18: Beispiel Slotabfragen


Slot	erfragte Werte
Typ(4)	geschäftliche Mobilnummer, geschäftliche Festnetznummer, private Mobilnummer oder private Festnetznummer?
Firma(6)	Kohlpharma, Möbel Martin, Globus, Sparkasse, Carglass oder Post
Nachname(6)	Meier, Bies, Schmidt, Bauer, Schuhmacher oder Schiller
Stadt(6)	Saarbrücken, Frankfurt, Köln, Berlin, Ingolstadt oder München

Die Personenprofile wurden auf die geänderten Slots angepasst. Abbildung 9 zeigt das neue Profil von Anke.

Anke Schumacher



geschäftliche Festnetznummer

 Mainzerstr. 23, 66121, Saarbrücken

 A.Schumacher86@gmx.de

 Immobiliengruppe

Abbildung 9: Personenprofil: Anke im 2. Versuch

Welche Slots pro Person abgefragt werden, zeigt Tabelle 19. Für jede Person werden zwei Slots abgefragt. Bei jedem der anzurufenden Kontakte wird nach dem Telefontyp gefragt, da für diesen Slot nur vier Füller möglich sind. Die Verteilung des zweiten Slots ist unterschiedlich.

Tabelle 19: Slotabfrage pro Person

Anke	Peter	Fritz	Kim
Typ	Typ	Typ	Typ
Nachname			Nachname
	Firma		
		Stadt	

6.2. Versuchsaufbau

Der Versuchsaufbau ist identisch zu dem aus Versuch 1. Tabelle 20 zeigt einen Überblick.

Tabelle 20: Übersicht Versuchsablauf

Vorrunde	1. Runde	2. Runde	3. Runde	4. Runde
Rennspiel	Rennspiel	Rennspiel	Rennspiel	
	Anruf Anke	Anruf Peter	Anruf Fritz	Anruf Kim

6.3. Versuchsdesign

Das Versuchsdesign wurde ebenfalls aus dem ersten Versuch übernommen. Ein Überblick der Strecken- und Strategieverteilung pro Gruppe ist in Tabelle 21 aufgelistet.

Tabelle 21: Strecken- und Strategieverteilung

Aufteilung	Strategie 1	Strategie 2	Strategie 3
1. Gruppe	Strecke A	Strecke B	Strecke C
2. Gruppe	Strecke B	Strecke C	Strecke A
3. Gruppe	Strecke C	Strecke A	Strecke B
4. Gruppe	keine Strecke	keine Strecke	keine Strecke

6.4. Control Panel

Das Control Panel aus Versuch 1 wurde mit anderen Sprachausgaben ausgestattet und es wurden weitere Schaltflächen für Strategie 3 hinzugefügt. (Abbildung 10)

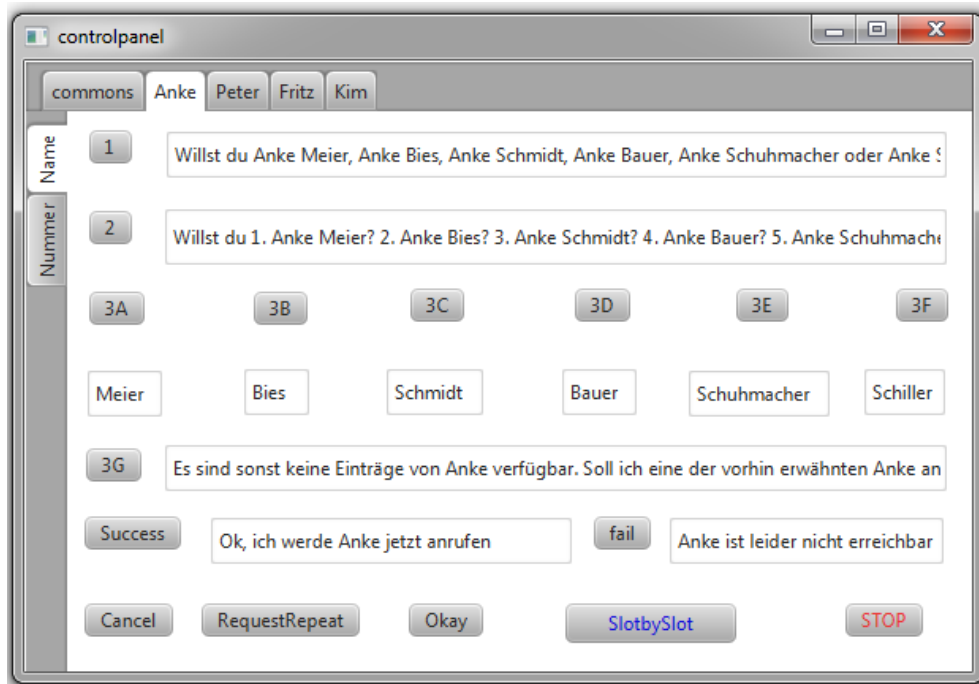


Abbildung 10: Controlpanel 2. Versuch

6.5. Versuchspersonen

Hier wurden ebenfalls 12 Muttersprachler getestet. Die Altersspanne der Probanden liegt zwischen 18 und 53. Jede Versuchsperson wurde zufällig einer Gruppe zugewiesen und hatte die selbe Aufgabe wie die Versuchsperson in Versuch 1:

1. Testrunde fahren
2. Fragebogen über Person ausfüllen (siehe 6.6.2 Unterkapitel Person)
3. Strecke A fahren + Anke anrufen
4. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen (siehe 6.6.2 Unterkapitel Fragebogen)
5. Strecke B fahren + Peter anrufen

6. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen
7. Strecke C fahren + Fritz anrufen
8. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen
9. Kim anrufen
10. Fragebogen über kognitive Belastung und letzten Dialog ausfüllen

Wie in Versuch 1 war die Aufgabe der Versuchspersonen ihre Eingaben deutlich über ein Tischmikrofon zu übermitteln. Dabei konnten sie sich die Personenprofile während des Dialoges auf einem Laptop ansehen.

6.6. Auswertung

Wie in Versuch 1 werden die Zeiten gemessen, die die Versuchsperson zum einen für das absolvieren der Strecke und zum anderen für das erfolgreiche Abschließen des Testszenarios benötigt (5.7.1 Gemessene Zeiten). Nach jeder Rennrunde wird die Versuchsperson ebenfalls einen Fragebogen ausfüllen, welcher sich auf die subjektiv wahrgenommene kognitive Belastung und auf Merkmale der Disambiguierungsstrategien bezieht (6.6.2 Fragebogen). Ebenfalls wird die Task Completion ausgewertet (6.6.3 Task Completion) und das Dialogverhalten untersucht (6.6.4 Dialogverhalten). Die erzielten Werte werden mit dem Tukey-Test auf statistische Signifikanz geprüft.

6.6.1. Gemessene Zeiten

Rennzeiten

In jeder Runde wurden die Rennzeiten bemessen. Mit diesen möchte man ermitteln, ob das Rennverhalten von einer Dialogstrategie beeinflusst wird. Die durch-

schnittlichen Rennzeiten für alle Strategien auf die einzelnen Strecken verteilt sind in Tabelle 22 gelistet.

Tabelle 22: Durchschnittszeiten Strategie pro Strecke

Rennzeiten	Strategie 1	Strategie 2	Strategie 3
Strecke A	81 sek	80,3 sek	74,25 sek
Strecke B	74,5 sek	84,25 sek	88 sek
Strecke C	75 sek	67,9 sek	65 sek

Durch diese Ergebnisse kann man keine Strategie bestimmen, mit der die Strecken am besten bzw. am schlechtesten gefahren wurden. Dies könnte daran liegen, dass einzelne Werte durch schlechtere bzw. bessere Spieler in den Gruppen verfälscht wurden. Tabelle 23 beinhaltet die durchschnittlichen Rennzeiten aller Strecken pro Strategie.

Tabelle 23: Durchschnittszeiten pro Strategie

Rennzeiten	Strategie 1	Strategie 2	Strategie 3
Durschnitt	76,83 sek	77,47 sek	76,73 sek

Die Zeiten weichen nur sehr gering voneinander ab und die Unterschiede sind zudem statistisch nicht signifikant. Diese Erkenntnis bekräftigt die in Versuch 1 getroffene Vermutung, dass die Rennzeit keinen Aufschluss darüber gibt, welche Strategie für die Autofahrt am Geeignetsten ist.

Dialogzeiten

Neben den Zeiten für das Rennspiel werden auch die Dialogzeiten berechnet. Die

Werte daraus werden dazu genutzt, um zu erforschen, mit welcher Strategie der kürzeste Dialog möglich ist. Weiter werden aufkommende Unterschiede im Dialogverhalten zwischen einer hoch belastenden und eine weniger belasteten Versuchsperson untersucht. Es werden alle Dialogzeiten aus den Runden mit Rennspiel gemessen und einmal für jede Strecke der Durchschnitt pro Strategie und einmal der gesamte Durchschnitt pro Strategie gebildet. Diese Werte lassen sich gegen die Durchschnittszeiten aus der Runde ohne Rennspiel vergleichen. Es werden allerdings nur die Zeiten gewertet, bei denen der Dialog eine maximale Task Completion erreicht hat. In Tabelle 24 sind die Durchschnittszeiten der gemessenen Runden aufgelistet.

Tabelle 24: Durchschnittsdialogzeiten 2. Versuch

Dialogzeiten	Strategie 1	Strategie 2	Strategie 3
Strecke A	25,76 sek	38,32 sek	33,98 sek
Strecke B	31,41 sek	40,2 sek	36,61 sek
Strecke C	29,59 sek	37,9 sek	28,29 sek
Strecke A - C	29,55 sek	38,54 sek	34,34 sek
ohne Strecke	24,12	34,35	30,44

Die zeitlichen Unterschiede der Disambiguierungsstrategien sind statistisch signifikant. An diesem Ergebnis sieht man, dass die erste Strategie den kürzesten Dialog ermöglicht und die zweite Strategie im Durchschnitt am Längsten dauert. Dies gilt sowohl für die Runden mit als auch für die Runden ohne Rennspiel. Der Vergleich der letzten beiden Zeilen der Tabelle macht deutlich, dass auch in diesem Versuch der Dialog ohne Rennspiel im Durchschnitt deutlich kürzer war, als der Dialog mit

Rennspiel. Dadurch wird vermutet, dass die Reaktionszeit bei geringer Belastung kleiner ist, als bei höherer Belastung.

6.6.2. Fragebogen

Zu Beginn des Versuchs wird derselbe Personenfragebogen wie in Versuch 1 ausgefüllt. Nach jeder Runde wird ebenfalls der Fragebogen, bestehend aus einem Teil des NASA-TLX Tests und einem Teil über die zuletzt gehörten Strategie, abgefragt.

Nasa-TLX

Dieser Teil des Fragebogens wird wie im ersten Versuch dazu genutzt um zu erforschen, bei welcher Strategie eine höhere Belastung empfunden wurde und ob es Unterschiede in der empfundenen Belastung in den Runden mit und ohne Rennspiel gibt. Die nachfolgende Tabelle zeigt die Ergebnisse jeder Frage pro Strategie.

Antwortenintervall	Strategien	Ergebnisse bestimmter Runden		
		1-4	1-3	4

Geistige Anforderung

1: gering 6: hoch	Strategie 1	2,31	2,67	1,25
	Strategie 2	2,31	2,58	1,5
	Strategie 3	2,56	2,83	1,75

Körperliche Anforderung

1: gering 6: hoch	Strategie 1	1,67	2,17	1
	Strategie 2	2,06	2,25	1,5
	Strategie 3	1,94	2,08	1,5

Zeitliche Anforderung

1: gering 6: hoch	Strategie 1	2	2,25	1,25
	Strategie 2	2,25	2,5	1,5
	Strategie 3	1,94	2,17	1,25

Leistung

1: gering 6: hoch	Strategie 1	5,13	4,83	6
	Strategie 2	4,88	4,67	5,5
	Strategie 3	4,56	4,08	6

Anstrengung

1: gering 6: hoch	Strategie 1	2,38	2,83	1
	Strategie 2	2,44	2,67	1,75
	Strategie 3	2,31	2,67	1,25

Frustration

1: gering 6: hoch	Strategie 1	1,94	2,25	1
	Strategie 2	1,94	2	1,75
	Strategie 3	2,31	2,58	1,5

Auch in diesem Versuch sind die Antworten der einzelnen Strategien rein statistisch nicht signifikant unterschiedlich und es ist kein eindeutiges Muster zu erkennen, welche Strategie am Wenigsten belastend ist. Beim Vergleich der Antworten von Runde 1-3 mit Runde 4 wird jedoch deutlich, dass bei allen Fragen die Runde mit Rennspiel als belastender gewertet wurde als die Runde ohne. Dieses Ergebnis bestätigt das Ergebnis aus Versuch 1 und bestärkt die Aussage, dass hier ein Unterschied in der Belastung empfunden worden ist.

Strategien

In diesem Fragebogen bewerten die Versuchspersonen die zuletzt verwendete Strategie anhand verschiedener Kategorien. Die Durchschnittsantworten sind in nachfolgenden Tabelle aufgelistet.

Antwortenintervall	Strategien	Ergebnisse bestimmter Runden		
		1-4	1-3	4

Der Dialog lenkte mich vom Rennspiel ab

1: kaum - 6: stark -	Strategie 1	-	3,5	-
	Strategie 2	-	2,92	-
	Strategie 3	-	3,17	-

Die Nachfragen erleichterten es mir, den Anruf korrekt aufzubauen

1: erleichterte es 6: erschwerte es	Strategie 1	2,44	2,58	2
	Strategie 2	2,25	2,33	2
	Strategie 3	2,38	2,5	2

Wussten Sie, wann das System Spracheingaben erwartete?

1: immer 6: nicht immer	Strategie 1	1,63	1,67	1,5
	Strategie 2	1,5	1,5	1,5
	Strategie 3	1,57	1,5	1,75

Wie gefiel Ihnen der Dialog insgesamt?

1: sehr gut 6: weniger gut	Strategie 1	2,94	2,83	3,25
	Strategie 2	2,44	2,42	2,5
	Strategie 3	2,38	2,3	2,5

Fiel es Ihnen einfacher, den Dialog ohne Rennspiel zu führen?

1: viel einfacher - 6: nicht einfacher -	Strategie 1	-	-	2,75
	Strategie 2	-	-	3
	Strategie 3	-	-	2,5

Welcher Anruf gefiel Ihnen insgesamt am besten?

Anruf bzw. Strategie auswählbar	Strategie 1	16,7%	-	-
	Strategie 2	33,3%	-	-
	Strategie 3	50%	-	-

Es kann keine Strategie identifiziert werden, die eindeutig am besten bewertet wurde. Die Fragen im Bezug auf Ablenkung, erleichtert Aufbau durch Nachfragen und Zeitpunkt der Spracheingaben wurden für Strategie 2 am besten bewertet. Die Frage "Wie gefiel Ihnen der Dialog insgesamt" erhielt für Strategie 3 den höchsten Wert. Dadurch gefiel den Versuchspersonen diese Dialogstrategie am besten. Diese Erkenntnis wird durch das Ergebnis der letzten Frage bestätigt, welche konkret nach der besten Strategie fragt. Auf diese Frage wurde am häufigsten mit Strategie 3 geantwortet. Parallel gilt dies für Strategie 1, welche den niedrigsten Wert erhielt und auch am seltensten bei der letzten Frage gewählt wurde.

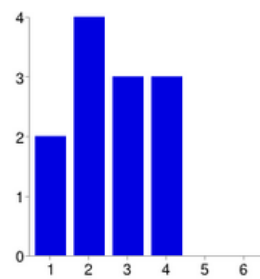
Die vorletzte Frage zeigt, dass es leichter fiel, den Dialog ohne Rennspiel zu führen und bestätigt damit auch hier das Ergebnis des NASA-TLX-Tests.

Person

In diesem Versuch werden die gleichen Informationen über die Versuchsperson zu Beginn des Fragebogens abgefragt. (vgl. Abbildung 7)

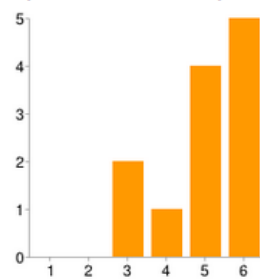
42% der Probanden sind in einer Altersgruppe von 18-29, 25 % in einer Altersgruppe von 30-41 und 33% in einer Altersgruppe von 42-53. Alle Versuchspersonen sind deutsche Muttersprachler. Eine Zusammenfassung der restlichen Antworten steht in Abbildung 11.

Haben Sie Erfahrung mit Dialogsystemen?



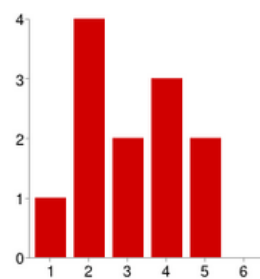
1	2	17 %
2	4	33 %
3	3	25 %
4	3	25 %
5	0	0 %
6	0	0 %

Spielen Sie oft Rennspiele?



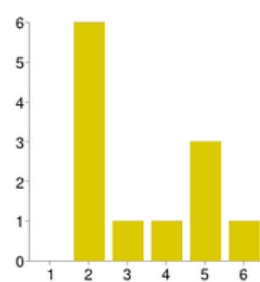
1	0	0 %
2	0	0 %
3	2	17 %
4	1	8 %
5	4	33 %
6	5	42 %

Wie technikaffin sind Sie?



1	1	8 %
2	4	33 %
3	2	17 %
4	3	25 %
5	2	17 %
6	0	0 %

Wie schwer fiel Ihnen die Einführungsrunde?



1	0	0 %
2	6	50 %
3	1	8 %
4	1	8 %
5	3	25 %
6	1	8 %

Abbildung 11: Fragebogen: Person

Das Ergebnis zeigt, dass 75% der Testpersonen keine bzw. wenig Erfahrung mit Dialogsystemen haben. Aus den Antworten lässt sich weiter schließen, dass 83% der Befragten selten Rennspiele spielen und die Einführungsrunde 58% schwer fiel. Diese Prozentangaben stimmen zufällig mit den Zahlen aus dem ersten Versuch überein. Dadurch kann ausgeschlossen werden, dass die Unterschiede der Resultate aus beiden Versuchen durch unterschiedliche Erfahrung der Versuchspersonen mit Dialogsystemen und Rennspielen zustande kommen.

6.6.3. Task Completion

Für jede Strategie wird ebenfalls die Task Completion ausgewertet, welche besagt, mit welchem Erfolg der Anruf ausgeführt wurde. Folgende Punktezahlen sind für jede Strategie möglich:

- 0 Punkte, wenn kein Slot richtig gefüllt wird
- 1 Punkt, wenn ein Slot richtig gefüllt wird
- 2 Punkte, wenn alle Slots richtig gefüllt werden

Zur Auswertung wird pro Strategie eine Durchschnittspunktzahl berechnet, welche in Tabelle 27 stehen.

Tabelle 27: Task Completion Versuch 2

Strategien	insgesamt	Runde 1-3	Runde 4
1. Strategie	1,88	1,83	2
2. Strategie	1,81	1,75	2
3. Strategie	1,56	1,42	2
Insgesamt		1,67	2

Hier zeigt sich klar, dass die Dialoge in Runde 4 ohne Fehler erfolgten und somit im Durchschnitt erfolgreicher waren als die Runden mit Rennspiel. Es fällt auf, dass die Strategie, die in diesem Versuch als am Beliebtesten ausgewertet wurde, am meisten Fehler aufweist. Dabei kommt erneut die Frage aus Versuch 1 auf, ob man hier eine Verbindung zwischen beliebtester Strategie und Task Completion ziehen kann und ob die Versuchspersonen bemerkten, dass sie die Slots falsch gefüllt haben. Grundsätzlich kann man an diesem Ergebnis sehen, dass insgesamt ein Anruf am häufigsten korrekt mit Strategie 1 und am seltensten korrekt mit Strategie 3 ausgeführt werden konnte.

6.6.4. Dialogverhalten

In diesem Versuch wurde mit denselben Antwortmöglichkeiten geantwortet wie in Versuch 1 (siehe Tabelle 16).

Die Häufigkeiten dieser Antworten pro Strategie aus Runde 1-3 sind in nachfolgender Tabelle aufgelistet.

Tabelle 28: Antwortenverteilung pro Strategie

Antwort- möglichkeiten	Strategie 1	Strategie 2	Strategie 3
Slots	100%	62,5%	0%
Position	0%	37,5%	0%
ja/nein	0%	0%	100 %

Diese Ergebnisse sind mit denen aus Versuch 1 zu vergleichen. Hier wurde bei hoher Belastung ebenfalls mit den gleichen Antwortmöglichkeiten geantwortet wie

bei niedriger Belastung, weshalb kein Unterschied im Dialogverhalten bei unterschiedlicher Belastung erkennbar ist.

6.7. Resultat

Aus den Resultaten aus Kapitel 6.6 wird die effizienteste Strategie ermittelt.

Durch die Dialogzeiten wird klar, dass Strategie 1 sowohl in den Runden mit, sowie ohne Rennspiel den kürzesten Dialog ermöglicht. Aus den Ergebnissen des NASA-TLX Fragebogens und den Rennzeiten kann kein Rückschluss auf die effizienteste Strategie gezogen werden. Der zweite Teil des Fragebogens zeigt, dass Strategie 2 weniger ablenkt, einfacher aufzubauen ist und man den Zeitpunkt der Spracheingaben einfacher erkennt. Strategie 3 gefiel den Versuchspersonen allerdings am besten. Die Task Completion zeigt jedoch, dass Strategie 3 insgesamt am Schlechtesten und Strategie 1 am besten abschnitt.

Durch dieses Ergebnis wird Strategie 1 aufgrund der kürzesten Dialogzeit und der besten Task Completion als effizienteste Strategie gewertet. Strategie 3 ist jedoch eindeutig die Beliebteste.

7. Ergebnisse

Durch die Ergebnisse der Versuche können die Strategien bezüglich verschiedener Komponenten auf Effizienz und Beliebtheit für eine automobiler Anwendung getestet werden.

Die Rennzeiten werden gemessen um zu überprüfen, ob eine Dialogstrategie das Rennverhalten stört. In keinem Versuch kann jedoch eine Strategie bestimmt werden, mit der die Rennstrecken am besten bzw. am schlechtesten gefahren werden. Die Unterschiede in den Zeiten der einzelnen Strategien sind zum Einen sehr gering und zum anderen statistisch nicht signifikant. Daher kann hier keine aussagekräftige Erkenntnis über eine effizienteste Strecke gezogen werden, da die Zeiten durch Zufall zustande gekommen sein könnten.

Neben den Rennzeiten wurde die Dialogzeit gemessen. Dies soll eine Erkenntnis darüber geben, mit welcher Strategie der kürzeste Dialog möglich ist. Weiter kann erforscht werden, ob Unterschiede in der Reaktionszeit zwischen einer kognitiv hoch belasteten und einer weniger belasteten Versuchsperson zu erkennen sind. In beiden Versuchen ist in allen Runden der kürzeste Dialog mit Strategie 1 möglich. Durch einen Vergleich der Dialogzeiten der Runden mit und ohne Rennspiel wird deutlich, dass die Dialogzeiten in den Runden ohne Rennspiel besser sind. Da zwischen den Runden mit und ohne Rennspiel kein unterschiedliches Dialogverhalten zu erkennen ist, kann man die Unterschiede nicht darauf zurückführen. Dadurch wird stark vermutet, dass die Probanden bei geringerer kognitiver Belastung eine schnellere Reaktionszeit aufweisen als bei hoher Belastung.

Durch den NASA-TLX Fragebogen nach jeder Runde wird analysiert, welche Strategie als am belastendsten empfunden wurde. Weiter soll aus den Antworten er-

mittelt werden, ob eine unterschiedliche Belastung zwischen den Runden mit und ohne Rennspiel gespürt wurde. Die Ergebnisse beider Versuche zeigen, dass keine Strecke eindeutig als am belastendsten gewertet wurde, zudem sind die Ergebnisse statistisch nicht signifikant. Der Vergleich der Antworten zwischen den Runden mit und ohne Rennspiel bestätigt die Aussage, dass die Probanden in den Runden mit Rennspiel in jedem Versuch eine höhere Belastung empfunden haben als in den Runden ohne.

Der Fragebogen über die getesteten Strategien zeigt wie die Versuchsperson die Strategien in Bezug auf verschiedene Kategorien bewerten. Im ersten Versuch, indem nur über zwei mögliche Slotfüller disambiguiert wurde, ist Strategie 1 am besten bewertet worden. Strategie 1 wurde in diesem Fragebogen auch als beliebteste Strategie gewertet. Am schlechtesten wurde Strategie 3 bewertet. Diese Ergebnisse sind nicht deckungsgleich mit den Ergebnissen aus Versuch 2. Bei diesem wurde zwar keine Strategie eindeutig am besten bewertet, jedoch wurde Strategie 3 als beliebteste Strategie und Strategie 1 als unbeliebteste Strategie gewertet. Hier ist ein deutlicher Unterschied zwischen beiden Versuchen zu erkennen. Prinzipiell zeigt das Ergebnis, dass die Beliebtheit der Strategien von der Länge der Disambiguierung abhängt. Das Resultat lässt vermuten, dass bei kurzer Disambiguierung eine möglichst schneller Dialog bevorzugt wird. Bei längerer Disambiguierung wird die Strategie bevorzugt, bei der man das Tempo durch das Annehmen bzw. Ablehnen des vorgeschlagenen Slotfüllers selbst bestimmen kann. Dies wird bei einer langen Sprachausgabe wahrscheinlich dadurch bevorzugt, da man nicht ununterbrochen konzentriert dem Dialog folgen muss. Dies fällt bei einem kurzen Dialog einfacher.

Um zu analysieren, welche Strategie am erfolgreichsten ist, wurde die Task Completion für beide Versuche ausgewertet. Die Ergebnisse beider Versuche fallen un-

terschiedlich aus. Im ersten Versuch erzielte Strategie 2 die höchste Task Completion und im zweiten Versuch Strategie 1. In beiden Versuchen war Strategie 3 am wenigsten erfolgreich. Dies deckt sich jedoch nicht mit dem Ergebnis, dass diese Strategie in Versuch 2 als beliebteste gewertet wurde. In einer zukünftigen Arbeit könnte die Frage geklärt werden, ob man hier einen Zusammenhang zwischen beliebtester und erfolgreichster Strategie ziehen bilden kann. Durch die Task Completion beider Versuche kommt man jedoch zu dem Resultat, dass die Dialoge mit niedriger Belastung erfolgreicher durchgeführt wurden.

Für beide Versuche wurde zusätzlich überprüft, mit welchen Antworten auf Rückfragen reagiert wurde. Dabei ist auffällig, dass die Verteilung der Antworten auf die einzelnen Strategien bei beiden Versuchen sehr ähnlich ist. In beiden Versuchen konnte zudem kein Unterschied im Dialogverhalten bei unterschiedlicher Belastung erkannt werden.

Zusammenfassend zeigen die Ergebnisse der beiden Versuche insgesamt, dass die Disambiguierungslänge bei der Bewertung der Strategien eine Rolle spielt. Bei der Formulierung einer Disambiguierung im Dialogsystem für das Auto, sollte der Dialogdesigner die Anzahl der möglichen Slotfüller beachten. Der erste Versuch zeigt deutlich, dass bei einer Disambiguierung über wenige Slotfüller eine Rückfrage in Strategie 1 am geeignetsten ist. Je nachdem ob auf Beliebtheit unter den Benutzern oder auf Effizienz wert gelegt wird, sollte sich bei längere Disambiguierung für Strategie 1 oder Strategie 3 entschieden werden.

8. Diskussion

In dieser Arbeit wurden zwei Versuche durchgeführt um drei Strategien auf Effizienz und Beliebtheit in einem Dialogsystem, das speziell im Rahmen einer automobilen Anwendung erstellt wird, zu testen. Um eine möglichst realistische Fahrsimulation zu erreichen, spielten die Versuchspersonen parallel zum Testszenario ein Rennspiel und füllten anschließend einen Fragebogen aus, der sich auf die subjektiv wahrgenommene kognitive Belastung und einer Bewertung der aktuellen Strategie bezog. Die Ergebnisse zeigen, dass sich bei einer Disambiguierung mit wenigen Slotfüllern die Strategie **Aggregierte Auswahl ohne Pause** (Strategie 1) besonders eignet. Besteht die Disambiguierung aus mehreren Slotfüllern zeigte sich die Strategie **Aggregierte Auswahl ohne Pause** (Strategie 1) als effizient und die **Sequentielle Auswahl** (Strategie 3) als beliebt unter den Benutzern.

In einer vierten Runde wurde das Testszenario in einer zufällig gewählten Strategie ohne Rennspiel durchgeführt. Es zeigte sich, dass die Versuchspersonen in dieser Runde eine geringere kognitive Belastung empfunden haben als in den Runden zuvor, sodass gesagt werden kann, dass das Rennspiel kognitiv belastend wirkte. Da die Dialogzeiten in der vierten Runde am Kürzesten sind kommt man hier zu dem Schluss, dass die Reaktionszeit bei geringer kognitiver Belastung besser ist als bei kognitiv belasteten Personen. Ein ähnliches Verhalten konnte in [Villing, 2009] festgestellt werden. In [Tsiakoulis et al, 2012] hat man herausgefunden, dass die Anzahl der Barge-Ins bei großer Belastung höher ist. Ein Unterschied im Dialogverhalten konnte in diesem Experiment nicht erkannt werden. Eine Erklärung hierfür ist, dass die Versuchspersonen die Möglichkeit zum Barge-In nicht erkannten. In [Tsiakoulis et al, 2012] wurde weiter erforscht, dass Dialogabläufe mit einfachen Spracheingaben wie **ja** oder **nein** bevorzugt werden. Das wurde durch das Ergebnis aus Versuch 2 bestätigt. Außerdem wird vermuten, dass die Versuchspersonen

sonen in Versuch 1 weniger belastet waren als die Versuchspersonen in Versuch 2. Die Ergebnisse des NASA-TLX Teils des Fragebogens zeigen, dass die eingeschätzte Belastung in Versuch 2 in vier von sechs Fragen im Durchschnitt (**Geistige Anforderung**, **Zeitliche Anforderung**, **Anstrengung** und **Frustration**) höher gewertet wurde. Dadurch wurde diese Vermutung bestärkt.

Die Task Completion beider Versuche zeigt, dass die Systeminteraktion ohne Rennspiel erfolgreicher ablief als die Systeminteraktion mit. Gleiche Ergebnisse wurden in [Tsiakoulis et al, 2012] festgestellt. Im ersten Versuch ist die Strategie **Aggregierte Auswahl mit Pause** (Strategie 2) am Erfolgreichsten und im zweiten Versuch die Strategie **Aggregierte Auswahl ohne Pause** (Strategie 1). An dieser Stelle stellt sich die Frage, ob die Task Completion in einer echten Systeminteraktion bei ähnlichem Testszenario genauso ausfällt. Die Task Completion kommt in dieser Studie dadurch zustande, in dem die Anzahl der richtig gefüllten Slots gewertet werden. Welche Slots gefüllt werden sollen, wird den Benutzern über ein Personenprofil angezeigt. In einer realen Situation entscheiden die Benutzer selbst, welche Slots gefüllt werden, was vermutlich zu einer geringeren Fehlerquote führt. Die Erkenntnis, dass im zweiten Versuch die Strategie **Sequentielle Auswahl** die meisten Fehler aufwies und parallel als beliebteste Strategie bewertet wurde, bestärkt diesen Verdacht.

Diese Studie fokussiert sich auf die geeignetste Disambiguierungsstrategie unter kognitiver Belastung. Dabei wurde die beliebteste Strategie bei keiner bzw. geringer kognitiver Belastung vernachlässigt. In einer zukünftigen Arbeit könnte man den Fokus umkehren und mehr Ergebnisse für das Testszenario ohne Rennspiel sammeln. Des Weiteren steht die Frage offen, ob ein Rennspiel eine reale Autofahrt simulieren kann. Die Rennzeiten konnten in diesem Experiment keinen Aufschluss

darauf geben, welche Strategie am wenigsten bzw. am meisten ablenkt. Dies könnte daran liegen, dass die Rennstrecken zu lang gewählt wurden. In einem zukünftigen Experiment könnte überprüft werden, ob das Experiment in einer realen Autofahrt ähnliche Resultate bringt und ob das Rennspiel eine Autofahrt annähernd simulieren kann.

Literatur

- [Ang et al., 2006] Chee Siang Ang, Panayiotis Zaphiris, Shumalai Mahmood: *Cognitive Load Issues in MMORPGs* (2006).
- [Jurafsky et al.] Daniel Jurafsky und James Martin: *Speech and Language Processing* Prentice Hall 2. Auflage (2008).
- [McCracken et al.] Daniel D. McCracken, Edwin D. Reilly: *Backus-Naur form (BNF)* Encyclopedia of Computer Science 129-131.
- [Minker et al., 2002] W. Minker, U. Haiber, P. Heisterkamp, S. Scheible: *intelligent dialog strategy for accessing infotainment applications in mobile environments* ISCA Tutorial and Research Workshop (ITRW) on Multi-Modal Dialogue in Mobile Environments, Irsee (Germany) (June 2002).
- [Mishra et al., 2004] R Mishra, E Shriberg, S Upson, J Chen, F Weng, S Peters, L Cavedon, J Niekrasz, H Cheng, und H Bratt. *A wizard of Oz framework for collecting spoken human-computer dialogs.* (2004)
- [Rogers et al.] Yvonne Rogers, Helen Sharp und Jennifer Preece: *Interaction Design: beyond human-computer interaction* John Wiley & Sons 3. Auflage (2011).
- [Tsiakoulis et al, 2012] P. Tsiakoulis, M. Henderson, B. Thomson, K. Yu, E. Tzirkel, S. Young: *The Effect of Cognitive Load on a Statistical Dialogue System* Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 74–78, Seoul, South Korea, (July 2012).

- [Villing, 2009] Jessica Villing: *Dialogue behaviour under high cognitive load* Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 322–325,(2009)
- [Yin et al., 2007] Bo Yin, Natalie Ruiz, Fang Chen, M. Asif Khawaja: *Automatic cognitive load detection from speech feature* in OZCHI '07: Proceedings of the 19th Australasian conference on Computer-Human Interaction 249-255.

A. Anhang

Abbildungsverzeichnis

1.	Funktionsweise der ODP S3 Plattform	3
2.	Personenprofil: Anke im 1. Versuch	13
3.	Rennspiel während des Versuchs	16
4.	Controlpanel Versuch 1	20
5.	Fragebogen: NASA-TLX	26
6.	Fragebogen: Dialogstrategien	30
7.	Fragebogen: Person	34
8.	Fragebogen: Person	36
9.	Personenprofil: Anke im 2. Versuch	44
10.	Controlpanel 2. Versuch	47
11.	Fragebogen: Person	56

Tabellenverzeichnis

1.	Interaktionsbeispiel Aggregierte Auswahl ohne Pause	9
2.	Interaktionsbeispiel Aggregierte Auswahl mit Pause (Zahl) . .	10
3.	Interaktionsbeispiel Aggregierte Auswahl mit Pause (Barge-In)	10
4.	Interaktionsbeispiel Sequentielle Auswahl	11
5.	Slotabfragen	14
6.	Slotabfrage pro Person	15
7.	Übersicht Versuchsablauf	17
8.	Strecken- und Strategieverteilung	18
9.	Anruf pro Strecke	19
10.	Durchschnittsrennzeiten jeder Strategie pro Strecke	22
11.	Durchschnittsrennzeiten pro Strategie	23
12.	Durchschnittsdialogzeiten	24
15.	Durchschnittliche Task Completion pro Strategie	38
16.	Antwortmöglichkeiten1	39
17.	Antwortenverteilung pro Strategie	40
18.	Slotabfragen	43
19.	Slotabfrage pro Person	45
20.	Übersicht Versuchsablauf	45
21.	Strecken- und Strategieverteilung	46
22.	Durchschnittszeiten Strategie pro Strecke	49
23.	Durchschnittszeiten pro Strategie	49
24.	Durchschnittsdialogzeiten 2. Versuch	50
27.	Task Completion Versuch 2	57
28.	Antwortenverteilung pro Strategie	58