UNTERSUCHUNG VON QUALITÄTSUNTERSCHIEDEN ZWISCHEN GESPROCHENER UND GESCHRIEBENER SPRACHE BEI INTERAKTION MIT EINEM CHATBOT

Marco Braune

Technische Universität Berlin
marco.braune@live.de

Kurzfassung: Der Beitrag untersucht Qualitätsunterschiede zwischen gesprochener und geschriebener Sprache bei der Interaktion mit einem Chatbot. Dazu wurde ein mittels *Dialogflow* erstellter Chatbot verwendet, der Informationen zu den Mensen des Universitätscampus Charlottenburg ausgibt. Der Chatbot wurde um ein Interface zur gesprochenen Interaktion erweitert, welches als Website implementiert ist. Die geschriebene Interaktion findet über den Messenger *Telegram* statt.

Es wurden zwei Evaluationen durchgeführt: Als erstes wurde eine Evaluation des Interfaces durchgeführt, um anschließend dessen Nutzbarkeit zu verbessern. Dazu wurde der Fragebogen der *System Usability Scale* und die Fragen des *Short Visual Aesthetics of Websites Inventory* verwendet. Daran haben 10 Personen teilgenommen.

Die zweite Evaluation diente der Untersuchung von Qualitätsunterschieden zwischen den beiden Interaktionsarten. Diese wurde als within-subjects Test im Feld durchgeführt. Eine Durchführung im Labor war aufgrund der COVID-Situation nicht möglich. Verwendet wurden die Fragebögen nach ITU-T Rec. P.851 2003, für die geschriebene Interaktion in angepasster Form, und der *MultiModal Quality Questionnaire*. Daran nahmen 31 Personen teil. Die Ergebnisse wurden in acht Komponenten analysiert und mittels gepaarten t-tests und Wilcoxon-Vorzeichen-Rang-Tests verglichen. Dabei ergab sich, dass die geschriebene Interaktion der gesprochenen Interaktion bevorzugt wird. Hauptsächliche Gründe dafür sind die Umgebungsunabhängigkeit und die Dauer der Interaktion.

1 Einführung

Systeme mit einer Interaktionsmöglichkeit zum Erfragen von Informationen werden immer verbreiteter. Zum einen gibt es Sprachassistenten, wie *Alexa* und *Siri*, die auf eine gesprochene Frage eine gesprochene Antwort liefern. Zum anderen gibt es auf Internetseiten oftmals die Möglichkeit über geschriebene Interaktion mit einem Chatbot zu kommunizieren. Während einige Systeme nur eine der beiden Interaktionsarten anbieten, gibt es hingegen auch viele Systeme, mit denen sowohl die Interaktion mittels Texteingabe, als auch mittels Spracheingabe möglich ist. Da es mit der Verbreitung dieser Systeme auch vermehrt Nutzer gibt, ist es wichtig die Qualität dieser bestimmen zu können. Während es bereits Verfahren zur Evaluierung der Qualität von Chatbots über gesprochene Interaktion gibt, gibt es hingegen kaum etablierte Verfahren zur Untersuchung der Qualität bei Interaktion über Texteingabe.

In dieser Arbeit geht es darum, die Qualitätsunterschiede zwischen gesprochener und geschriebener Sprache bei der Interaktion mit einem Chatbot zu untersuchen. Verwendet wird dabei der Chatbot „Lotti“, ein Dialog-System, das Informationen zu den Mensen auf dem Universitätscampus Charlottenburg ausgibt. Bisher ließ sich damit nur über geschriebenen Text interagieren, weshalb der Chatbot zunächst um ein Sprachinterface erweitert wurde. Im Folgenden werden zunächst die wichtigsten Unterschiede zwischen der geschriebenen und gesprochenen Interaktion genannt. Dann wird die Funktionalität des Chatbots „Lotti“, vorge-

stellt und die Evaluation des Sprachinterfaces beschrieben. Im Abschnitt zur Evaluation der Qualitätsunterschiede zwischen den Interaktionsarten wird sowohl das Vorgehen beschrieben als auch die Ergebnisse präsentiert, welche dann in der Diskussion ausgewertet werden. Auch wird die Möglichkeit, die Methodik zukünftig anzupassen, diskutiert. Abschließend werden Aspekte betrachtet, die sich aus der Arbeit schließen lassen und für weitere Untersuchungen interessant wären.

2 Unterschiede zwischen Sprach und Textinteraktion

Bei der im Folgenden untersuchten Interaktion wird entweder Text oder Sprache als Kommunikationsweg verwendet. Im technischen Aufbau unterscheidet sich der Chatbot für die jeweilige Interaktionsart in den verschiedenen Komponenten. Beide Systeme bestehen aus der natürlichen Spracherkennung, einem Dialog Manager verbunden mit einer Datenbank und einer Antwort Generierung. Systeme, die gesprochene Interaktion unterstützen (Sprach-Dialog-Systeme), verfügen zusätzlich noch über einen Spracherkenner und einen Sprachsynthetisierer [4].

Abgesehen von der beschriebenen Architektur des Systems, gibt es insbesondere auf der Seite des Nutzers zwischen den beiden Interaktionsarten wesentliche Unterschiede: Bei der Interaktion über Text hat der Nutzer die Möglichkeit, die bereits erhaltene Information auch später noch nachzulesen. Bei der Interaktion über Sprache muss der Nutzer die Information bei Bedarf erneut erfragen. Außerdem beschreibt Zoltan-Ford [9], dass insbesondere die Interaktion über Text mehr Aufwand erfordert als die Sprachinteraktion. Dies kann dazu führen, dass Nutzer mehr Interaktionen beim Sprechen als beim Schreiben durchführen. Zudem können sich die Nutzer beim Schreiben auch mehr Zeit für die Anpassung der Wortwahl ihrer Anfrage nehmen und diese vor dem Senden an das System überarbeiten, während eine gesprochene Aussage oftmals spontaner geschieht.

3 Chatbot „Lotti“

Der hier zur Untersuchung verwendete Chatbot „Lotti“ wurde von J. Zollner mittels *Dialogflow* erstellt [8]. Anhand unterschiedlicher in *Dialogflow* erstellter Intents wird eine passende Antwort auf eine Anfrage gegeben. Je nach Intent werden die Anfragen über einen Webhook weiterverarbeitet. Der Webhook ist eine Serverdatei, die als Schnittstelle zu der *OpenMensa* Datenbank dient. Die jeweilige Abfrage wird darüber an die Datenbank übermittelt, um die aktuelle Information zu den Mensen des Universitätscampus Charlottenburg abzurufen. Mögliche Informationen, die erfragt werden können, sind die Speisepläne des jeweils genannten Tages, Filterung bezüglich der Speisen nach Allergenen oder Vorlieben, Erläuterungen der verwendeten Symbolik in den Speiseplänen und Standortdetails, Websiteadressen und Öffnungszeiten der einzelnen Mensen. Außerdem kann auch diverser Smalltalk mit "Lotti" geführt werden, wie zum Beispiel eine Fragestellung nach dem Wohlbefinden.

Die geschriebene Interaktion mit dem Chatbot findet über den Messenger *Telegram* statt. Sowohl die Eingabe des Nutzers als auch die Ausgabe des Chatbots ist dabei in Textform. Um auch gesprochene Interaktion verwenden zu können, wurde im Rahmen dieser Arbeit noch ein Sprachinterface implementiert, welches als Website aufgerufen wird. Die Eingabe erfolgt ausschließlich über ein Mikrofon, die Ausgabe je nach Intent sprachlich über einen Audioausgang oder für ausführlichere Informationen wie den Speiseplan in Textform auf der Webpage. Bei der Ausgabe von Adressen erfolgt die Ausgabe sowohl sprachlich als auch textlich.

Das Sprachinterface wurde zunächst alleinstehend evaluiert, um die Funktionalität für die eigentliche Untersuchung mit Sicherheit zu gewährleisten. Es wurden 10 Personen befragt. Dabei wurde der Fragebogen zur *System Usability Scale* (*SUS*) [2] verwendet und die Fragen des *Short Visual Aesthetics of Websites Inventory* (*VisAWI-S*) [5]. Das Sprachinterface erzielte

einen *SUS*-Score von 75,25 und ist damit nach Brooke [2] als „Gut“ zu bewerten (siehe Abbildung 1). Der Ästhetik-Faktor des *VisAWI-S* ergab einen Wert von 3,58 und liegt dabei im mittigen Bereich. (Es wurde, wie auch beim SUS, eine fünfstufige anstatt einer siebenstufigen Likert-Skala verwendet, um die Antwortskala innerhalb der einen Umfrage konsistent zu halten.) Anschließend wurden noch kleinere Änderungen vorgenommen, um auf genannte Kritikpunkte einzugehen und das Interface insgesamt zu verbessern.

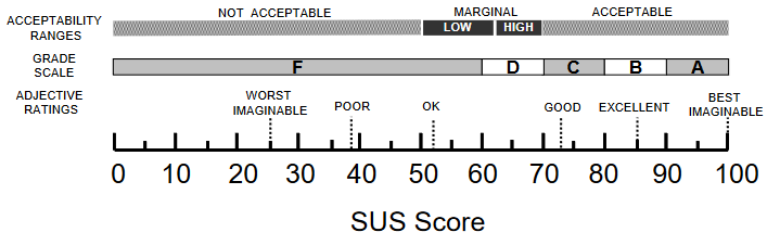

Abbildung 1 - Ein Vergleich der Adjektivbewertungen, Akzeptanzwerte und Schulnotenskala auf den durchschnittlichen SUS-Wert (Quelle: Brooke [2])

4 Evaluation

Zur Untersuchung der Qualitätsunterschiede zwischen gesprochener und geschriebener Sprache bei der Interaktion mit einem Chatbot wurde eine Umfrage mit 31 Teilnehmern durchgeführt. Verwendet wurden der Fragebogen der ITU-T Rec. P.851 2003 [3] und der *Multimodal Quality Questionnaire* (*MMQQ*) [7]. Der ITU-T Fragebogen wurde so angepasst, dass dieser nicht ausschließlich für gesprochene, sondern auch für geschriebene Interaktion verwendbar ist. Der *MMQQ* wurde verwendet, um auch die Multimodalität des Sprachinterfaces zu evaluieren.

4.1 Durchführung

Für die Evaluierung wurde ein within-subjects Testdesign angewandt. Die Durchführung erfolgte aufgrund der COVID-Situation als Feldtest. Es sollte ausschließlich über einen Computer und nicht über ein Mobilgerät teilgenommen werden. Jeder Teilnehmer erhielt je Interaktionsart drei unterschiedliche Aufgaben. Die Reihenfolge der Aufgaben wurde dabei in Anlehnung an die Erstellung des lateinischen Quadrats [1] randomisiert. Es gab folgende Aufgaben:

1. Finde heraus welche Mensen der Chatbot kennt
2. Führe Smalltalk mit dem Chatbot, indem Du nach dem Alter fragst.
3. Lass Dir genauere Infos zu einer Mensa Deiner Wahl anzeigen. (Bsp.: Skyline Mensa)
4. Lass Dir den Speiseplan von heute in einer Mensa Deiner Wahl anzeigen (Bsp.: Skyline Mensa) und anschließend von morgen.
5. Lass Dir den Speiseplan von einem bestimmten Wochentag in einer Mensa Deiner Wahl anzeigen (Bsp.: Skyline Mensa) und anschließend nur die veganen Speisen davon.
6. Lass Dir nur die mit der roten Ampel gekennzeichneten Speisen in einer Mensa Deiner Wahl anzeigen (Bsp.: Skyline Mensa) und anschließend alle Speisen in dieser Mensa.

4.2 Ergebnisse

Etwa 48% der Teilnehmer waren Angestellte oder Studierende der TU Berlin oder der UdK Berlin und dadurch mit dem Campus und den Mensen vertraut. Da in den Aufgabenstellungen eine Mensa vorgeschlagen wurde, sollte dies dennoch keinen Einfluss der Bewertung von

Teilnehmern haben, die nicht mit den Mensen vertraut waren. Bis auf einen beherrschten alle die deutsche Sprache auf Muttersprachenniveau, weshalb es keine Probleme bei der Verständigung aufgrund der Sprache gegeben haben sollte. Auch nutzte der Großteil nicht häufig Chatbots.

Der Gesamteindruck wurde auf einer Likert-Skala von „Schlecht“, das dem Wert -2 entspricht, bis „Ausgezeichnet“, das dem Wert 2 entspricht, bewertet. Der Gesamteindruck der Sprachinteraktion des Systems wurde durchschnittlich mit einem Wert von 0,39 bewertet, der Gesamteindruck der geschriebenen Interkation hingegen mit 0,94. Auch in der expliziten Frage nach der Präferenz gaben 81% an, die Interaktion über *Telegram* zu bevorzugen und nur 6% die Interaktion über das Sprachinterface. 13% empfanden beide Interaktionsarten als gleich gut.

Die Multimodalität wurde durchschnittlich bei keiner Aussage negativ bewertet. Lediglich die Bewertung, ob die Interaktion „eintönig“ oder „abwechslungsreich“ sei, liegt direkt in der Mitte. Damit sprechen die Ergebnisse aber dafür, dass sich die Multimodalität nicht negativ auf die Bewertung der Interaktion auswirkt.

Zur Auswertung des ITU-T P.851 Fragebogens [3], werden die Aussagen in einzelnen Komponenten analysiert, wie es von Möller et al. [6] beschrieben wird. Der Wert der einzelnen Komponente ergibt sich aus den Mittelwerten der jeweiligen Bewertungen der Fragen. Die einzelnen Bewertungen wurden ebenfalls, wie bei der Gesamtbewertung, auf Werte von -2 bis 2 skaliert. Die mit einem "*" (Asterisk) markierten Aussagen, sind negative Aussagen und wurden so skaliert, dass ein hoher Wert ebenfalls für eine positive Bewertung und ein niedriger Wert für eine negative Bewertung steht. Die Ergebnisse dieser Analyse sind der Tabelle 1 zu entnehmen.

Tabelle 1 – Analyse in Komponenten zur Auswertung des Fragebogens nach ITU-T P.851 (2003)

Komponente / Aussage	Mittelwert **(Sprache)**	Mittelwert **(Text)**
C1 – Acceptability (dt. Akzeptanz)	0,34	0,95
Ich würde die Funktionen lieber auf andere Weise bedienen.*	- 0,32	0,39
Ich würde den Chatbot in Zukunft wieder benutzen.	0,23	0,71
Ich konnte die Interaktion wie gewünscht lenken.	0,55	1,16
Ich bin insgesamt mit dem Chatbot zufrieden.	0,48	0,97
Der Chatbot ist nicht hilfreich zur Bedienung der Funktionen.*	0,39	0,81
Die Interaktion mit dem Chatbot war angenehm.	0,71	1,32
Mit dem Chatbot lassen sich die Funktionen effizient bedienen	0,32	0,94
Die Bedienung der Funktionen durch gesprochene / geschriebene Sprache war komfortabel.	0,39	1,29
C2 – Cognitive demand (dt. Kognitive Anforderung)	0,44	1,26
Ich musste mich sehr auf die Interaktion mit dem Chatbot konzentrieren.*	-0,03	1,19
Ich konnte den roten Faden während der Interaktion leicht verlieren.*	0,61	1,16
Ich fühlte mich entspannt.	0,74	1,42
C3 – Task efficiency (dt. Aufgabeneffizienz)	0,24	0,55
Der Chatbot reagiert nicht immer wie erwartet.*	-0,26	0,10
Die vom Chatbot gelieferten Informationen waren klar und deutlich.	0,90	1,58
Der Chatbot tat nicht immer das, was ich wollte.*	0,07	-0,03

C4 – System errors (dt. Systemfehler)	0,58	0,86
Der Chatbot machte viele Fehler.*	0,68	0,87
Der Chatbot ist unzuverlässig.*	0,48	0,84
C5 – Ease of use (dt. Benutzerfreundlichkeit)	0,98	1,44
Ich musste mich konzentrieren, um den Chatbot akustisch zu verstehen / die Nachrichten des Chatbots zu lesen.*	0,87	1,29
Die Benutzung des Chatbots lässt sich leicht erlernen.	1,10	1,58
C6 – Cooperativity (dt. Kooperativität)	0,84	1,03
Der Chatbot verhielt sich kooperativ.	0,84	1,03
C7 – Naturalness of system & symmetry of the dialogue (dt. Natürlichkeit des Systems & Ausgewogenheit des Dialogs)	0,18	0,84
Die Anteile in der Interaktion waren zwischen mir und dem Chatbot gleich verteilt.	0,45	0,74
Die Stimme / Die Antwort des Chatbots klang natürlich.	-0,10	0,94
C8 – Speed of interaction (dt. Geschwindigkeit der Interaktion)	0,29	1,58
Der Chatbot reagierte zu langsam.*	0,29	1,58

Zusätzlich wurde in SPSS auf Signifikanz der Unterschiede zwischen den beiden Interaktionsarten untersucht. Dazu wurden sowohl gepaarte t-tests als auch Wilcoxon-Vorzeichen-Rang-Tests durchgeführt. Aus den Tests ergibt sich, dass die Bewertungen der Komponenten „Task efficiency“, „System errors“ und „Cooperativity“ mit einem Signifikanzwert von $p>0,05$ nicht signifikant unterschiedlich sind. Die statistischen Tests für die Bewertungen der weiteren fünf Komponenten ergaben, dass bei diesen ein signifikanter Unterschied vorliegt.

Des Weiteren wurden folgende Interaktionsparameter untersucht: Dauer der gesamten Interaktion, Zeit pro Interaktion, Anzahl der gesamten Interaktionen, Anzahl der aufgabenbezogenen Interaktionen, Anzahl der explorativen Interaktionen. Die Unterschiede der verschiedenen Interaktionsanzahlen sind nicht signifikant. Die Dauer der Interaktion, die bei der Textinteraktion 5 Minuten und bei der Sprache 6,22 Minuten beträgt, ist jedoch zu beachten, da einige Teilnehmer durchaus die Dauer der Sprachinteraktion in einem Freitextfeld kritisiert haben. Außerdem wurde die „IntentDetectionConfidence“ der einzelnen Interaktionsschritte ausgelesen und analysiert. Mit einem Signifikanzwert von $p=0,05$ ist der Unterschied zwischen den beiden Interaktionsarten gerade noch signifikant. Dies deutet darauf hin, dass die Intents über Texteingabe zuverlässiger erkannt werden. Dennoch ist dies keine Garantie, dass es die Intents sind, die dem Nutzer die korrekte Antwort liefern.

5 Diskussion

Um die Ergebnisse zu verstehen, werden diese im Folgenden diskutiert. Außerdem wird diskutiert, was sich aus der durchgeführten Evaluation für zukünftige Evaluationen schlussfolgern lässt.

5.1 Auswertung

Die geschriebene Interaktion wird der gesprochenen Interaktion bevorzugt. Die dafür verantwortlichen Faktoren werden in der Komponentenanalyse und auch bei der Methodik deutlich.

5.1.1 Auswertung der Komponenten

Drei der acht Komponenten - C3, C4 und C6 - wurden als nicht signifikant unterschiedlich ausgewertet. Diese Komponenten beziehen sich insbesondere auf das System an sich, welches bei beiden grundlegend gleich ist, anstatt die jeweilige Interaktionsart. Deshalb werden die Aussagen dieser Komponenten auch bei der Gesamtbewertung und Präferenz weniger Ein-

fluss gehabt haben. Aufgrund dieser Annahme werden im Folgenden besonders die signifikant unterschiedlichen Kategorien genauer analysiert.

Die erste Komponente stellt die Akzeptanz der beiden Interaktionen gegenüber. Neben der allgemeinen Interaktion ist auch das gesamte System zu betrachten. *Telegram* oder ein ähnlicher Messenger ist vielen Menschen - damit wahrscheinlich auch vielen Teilnehmern - vertraut, weshalb dies einige Aussagen positiv beeinflussen kann. Die Multimodalität des Sprachinterfaces wurde zwar größtenteils ebenfalls positiv bewertet, doch haben einige auch angegeben, dass sie sich es in einem anderen Format wünschten, wie beispielsweise einer App. Eine App wäre leichter zugänglich und auch vergleichbarer mit *Telegram*. In der reinen Interface Evaluation wurde angegeben, dass eine automatische Spracherkennung von Vorteil wäre, statt per Knopfdruck. Das heißt, die Nutzung des Interfaces an sich, kann sich auch direkt auf die Akzeptanz auswirken. Es gibt jedoch auch Aussagen wie „Die Bedienung der Funktionen durch gesprochene / geschriebene Sprache war komfortabel“, bei der sich ebenfalls deutlich abzeichnet, dass die Interaktion über Text an sich bevorzugt wird.

Im Gegensatz zu der Aussage von Zoltan-Ford [9] zu Beginn der Arbeit, dass die Interaktion über Sprache leichter fällt, widerspricht der Wert der Komponente bezüglich der kognitiven Anforderung. Hier geben die Teilnehmer an, die Anforderung sei geringer bei der geschriebenen Interaktion. Gründe sind dafür, dass zum einen die Antworten des Chatbots sehr umfangreich sind. Ebenfalls zu Beginn wurde erläutert, dass zum anderen Antworten bei geschriebener Interaktion nachgelesen werden können. Bei Sprache hingegen, müssen die Nutzer aufmerksam zuhören und desto ausführlicher die Antwort, desto fordernder ist dies.

Es wurde einmal angegeben, dass die Sprachausgabe leise war. Das heißt, der Nutzer muss zusätzlich sein Gerät anpassen und die Lautstärke kann sich je nach Gerät unterscheiden, was die Konzentration erhöht, den Chatbot akustisch zu verstehen. Text hingegen ist hauptsächlich auf die gleiche Weise verfügbar ist. Lediglich Bildschirmgröße und Auflösung spielen dabei eine Rolle. Dies sind Faktoren, die die Bewertung der Benutzerfreundlichkeit der Systeme beeinflussen. Jedoch fiel diese Komponente bei beiden Interaktionsarten besonders positiv aus. Dennoch fiel die Erlernung bei *Telegram* noch besser aus, was wieder darauf schließen lässt, dass der etablierte Messenger ein Vorteil in der Nutzung bietet. Doch auch das Sprachinterface verursachte keine großartigen Probleme bei der Nutzung.

Ein weiterer signifikanter Unterschied liegt bei der Natürlichkeit des Systems und der Ausgewogenheit des Dialogs. Die Natürlichkeit des Systems der geschriebenen Interaktion wird sehr positiv bewertet, die der gesprochenen Interaktion jedoch deutlich schlechter. Geschriebener Text wird in vielen Medien, sowohl digitale als auch gedruckte, zur Informationsausgabe genutzt. Von daher ist der Mensch dies gewohnt und empfindet es nicht als äußerst unnatürlich. Sprache ist der Mensch jedoch größtenteils von anderen Menschen gewohnt. Insbesondere bei den Teilnehmern dieser Evaluation ist die Nutzung von anderen Chatbots eher selten. Somit wird synthetische Sprache als deutlich weniger natürlich bewertet. Die Gleichverteilung der Interaktionsanteile ist erneut auf die Ausführlichkeit der Antworten zurückzuführen. Beide Systeme sind identisch und liefern dieselben Antworten, von daher sollte auch die Gleichverteilung in etwa identisch sein. Doch der Nutzer nimmt die Menge der Interaktion, die vom System ausgeht, möglicherweise unterschiedlich wahr. Denn eine einzelne Textnachricht zu erhalten, wird dem Nutzer als deutlich geringerer Interaktionsanteil des Systems vorkommen, als sich dieselbe Information als Sprachausgabe über mehrere Sekunden lang anzuhören.

Bei der Bewertung der Geschwindigkeit der Interaktion, sind die Interaktionsparameter heranzuziehen. Die Unterschiede dieser sind zwar nicht signifikant, aber dennoch eine Grundlage für die Unterschiede in der Bewertung. Gerade bei der Dauer der Interaktionen sind Unterschiede festzustellen: Die gesprochene Interaktion dauerte durchschnittlich mehr als sechs

Sekunden länger als die geschriebene Interaktion. Außerdem wurde bei der gesprochenen Interaktion durchschnittlich etwa eine Interaktion mehr benötigt, um die gestellten Aufgaben zu bearbeiten. Auch im Freitextfeld wurde oftmals die Reaktionszeit genannt, die bei der Textinteraktion schneller war. Während sich die Interaktionsanzahl direkt auf die Art der Interaktion bezieht, liegt der Grund für die Reaktionszeit, und damit gegebenenfalls für die gesamte Dauer, bei der Performance der Systeme. Da das Sprachinterface über die Google APIs noch Text-to-Speech und Speech-to-Text Transkriptionen durchführt, nimmt dies mehr Zeit in Anspruch, als direkt die geschriebene Anfrage zu verarbeiten, was sich bei dem Nutzer als insgesamt beanspruchte Zeit widerspiegelt.

5.1.2 Methodik

Wie beschrieben, fand die Evaluation als Feldtest am Computer statt. Auf der einen Seite ist ein Feldtest geeignet, da der Chatbot in jeder möglichen Situation genutzt werden soll. Auf der anderen Seite kann so die Durchführung nicht genau kontrolliert werden, was zu verfälschten Ergebnissen führen kann.

Eben gerade um eine unkomplizierte Informationsausgabe zu den Mensen zu bieten, die jederzeit möglich sein soll, wäre es eine Option, den Chatbot auch mobil verfügbar zu machen. Einige hätten die mobile Nutzung sogar präferiert. Ein Smartphone hat der Großteil von Studenten bei sich und kann bei jeder Gelegenheit genutzt werden. Extra einen Computer auf dem Campus zu starten, kann aufwendiger sein, insbesondere wenn der Nutzer schon auf dem Weg zu einer Mensa ist.

5.2 Implikationen für die Evaluationsmethodik

Da der verwendete Fragebogen der ITU-T Rec. P.851 [3] für die gesprochene Interaktion mit Sprach-Dialog-Systemen entwickelt, hier aber auch für die geschriebene Interaktion angepasst und verwendet wurde, lassen sich daraus Schlussfolgerungen ziehen, inwieweit dieser für die Qualitätsuntersuchungen bei der geschriebenen Interaktion erweitert werden könnte.

Zu Beginn der Arbeit und in den Antworten der Evaluation wurde genannt, dass Nachrichten erneut nachgelesen werden können. Interessant wäre es zu erfragen, wie oft diese Möglichkeit genutzt wird, um herauszufinden, ob es angenehmer und gegebenenfalls schneller ist, die Anfrage erneut zu stellen oder die Information im bisherigen Verlauf nachzulesen. Weiterführend könnte dabei auch erfragt werden, bis zu wie vielen vergangenen Nachrichten zurückgeschaut wird, um die Information wiederzufinden, bevor die Frage erneut gestellt wird. Somit gäbe es einen Vergleich, um zu ermitteln, inwiefern diese Option einen Vorteil gegenüber der Sprachinteraktion bietet. Des Weiteren wird zwar nach der Klarheit und Deutlichkeit der Aussage gefragt, doch bei Textnachrichten trägt auch das Layout des Textes an sich zu der Aufnahme der Information bei. Dies war besonders auffällig bei der Evaluation des Interfaces, bei der einige die textliche Ausgabe kritisiert hatten. Aspekte des Textlayouts sind Strukturierung, Schriftart, Schriftgröße, Formatierung der Schrift und verwendete Symbole, wie hier beispielsweise Emojis.

Auch gibt es Aspekte, die allgemein bei der Durchführung zusätzlich beachtet werden können: Feldtests können zwar durchaus geeignet für ein solches System sein, doch da die Umgebung für viele ein relevanter Aspekt ist, sollte explizit erfragt werden, in welcher Umgebung die Evaluation durchgeführt wurde. Zusätzlich wäre es sinnvoll, in zuvor definierten Szenarien zu testen. Mit diesen beiden Ergänzungen könnten sowohl die Bewertung des einzelnen Systems, als auch die Bewertung beider Systeme je nach Umgebung verglichen werden. Außerdem könnten zukünftig die Geräte erfragt werden. Sowohl die Eingabe als auch die Ausgabegeräte. Zwar sollte bei dieser Evaluation jeder am Computer teilnehmen, doch kann dabei noch zwischen Laptop und PC unterschieden werden. Bei den Ausgabegräten können

integrierte oder externe Lautsprecherboxen verwendet werden. Ebenso bei der Eingabe, entweder intergiertes Mikrofon oder ein Headset. Wenn die Umfrage auf Mobilgeräte ausgeweitet werden würde, kommen noch mehr Optionen bei den Geräten hinzu, aber auch ob Autokorrektur bei der Eingabe verwendet wurde. Bei der Analyse der Interaktionsparameter wurde festgestellt, dass einige Teilnehmer auch nach der Durchführung der Evaluation den Chatbot hin und wieder noch genutzt haben. Deshalb wäre es eine Idee, die Methodik weiterzuführen, indem der Chatbot im Alltag über einen bestimmten Zeitraum hinweg genutzt werden würde, anstatt, dass vordefinierte Szenarien und Aufgaben gestellt werden. Die Bewertung könnte dabei entweder in festgelegten Intervallen, wie nach Interaktionen, oder im Anschluss an den Zeitraum erfolgen. Wenn ein Zeitraum über mehrere Wochen bestimmt werden würde, könnte dem Nutzer auch die Möglichkeit gegeben werden, selbst auswählen zu können, welchen Chatbot er nutzen möchte, anstatt in einer Frage die Präferenz angeben zu müssen.

6 Fazit

Auch wenn die Anzahl der Teilnehmer mit 31 relativ gering ist und der Anwendungsbereich des verwendeten Chatbots begrenzt, gibt diese Untersuchung dennoch bereits Hinweise darauf, nach welchen Kriterien die geschriebene und gesprochene Interaktion bewertet werden und was Personen bei der Nutzung von Chatbots wichtig ist. Es lässt außerdem darüber nachdenken, ob sich diese Kriterien, die für Nutzer solcher Systeme relevant sind, noch spezifischer erfragen lassen.

Um eine allgemeine Aussage über Qualitätsunterschiede zwischen geschriebener und gesprochener Interaktion machen zu können, sollte die Evaluation in noch größerem Umfang durchgeführt werden.

Literatur

[1] Bortz, Jürgen und Christof Schuster: *Statistik für Human- und Sozialwissenschaftler*. Springer, 7. Auflage, 2010, ISBN 978-3-642-12769-4.

[2] Brooke, John: *SUS: A 'Quick and Dirty' Usability Scale*. In: Jordan, Patrick W., B. Thomas, Ian Lyall McClelland und Bernard Weerdmeester (Herausgeber): *Usability Evaluation In Industry*, Kapitel 21, Seiten 189–194. CRC Press, 1. Auflage, 1996.

[3] ITU-T Rec. P.851: *Subjective quality evaluation of telephone services based on spoken dialogue systems.* International Telecomunnication Union, Geneva, Switzerland, 2003.

[4] Lamel, Lori, Wolfgang Minker und Patrick Paroubek: *Towards best practice in the development and evaluation of speech recognition components of a spoken language dialog system*. Natural Language Engineering, 6(3-4):305–322, 2000. https://doi.org/10.1017/S1351324900002515.

[5] Moshagen, M. & Thielsch, M. T.: *A short version of the visual aesthetics of websites inventory. Behaviour & Information Technology*, 32(12):1305– 1311, 2013. https://doi.org/10.1080/0144929X.2012.694910. 57

[6] Möller, Sebastian, Paula Smeele, Heleen Boland und Jan Krebber: *Evaluating spoken dialogue systems according to de-facto standards: A case study*. Computer Speech & Language, 21(1):26 – 53, 2007, ISSN 0885-2308. https://doi.org/10.1016/j.csl.2005.11.003.

[7] Wechsung, Ina: *An Evaluation Framework for Multimodal Interaction - Determining Quality Aspects and Modality Choice*. Springer International Publishing, 2014, ISBN 978-3-319-03810-0.

[8] Zollner, Julia: *Konzeption und Implementierung eines Mensa-Chatbots für den Campus Charlottenburg*. Bachelorarbeit, Technische Universität Berlin, 2019.

[9] Zoltan-Ford, Elizabeth: *How to get people to say and type what computers can understand. International Journal of Man-Machine Studies*, 34(4):527 – 547, 1991, ISSN 0020-7373. https://doi.org/10.1016/0020- 7373(91)90034-5.