Inauguraldissertation zur Erlangung des Doktorgrades der Philosophie der Ludwig-Maximilians-Universität München

Institut für Phonetik und Sprachverarbeitung

Kontextabhängiges Prompt- und Flowdesign in Sprachassistenzsystemen

vorgelegt von

Anna-Maria Meck

aus

München

2024

Erstgutachter: PD Dr. Christoph Draxler, Ludwig-Maximilians-Universität München
Zweitgutachter: Prof. Dr. Albrecht Schmidt, Ludwig-Maximilians-Universität München

Tag der mündlichen Prüfung: 20.02.2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF INCLUDED PUBLICATIONS

Paper I

Anna-Maria Meck, Christoph Draxler, and Thurid Vogt. 2022. A Question of Fidelity: Comparing Different User Testing Methods for Evaluating In-Car Prompts. In Proceedings of the 4th Conference on Conversational User Interfaces (CUI ’22). Association for Computing Machinery, New York, NY, USA, Article 31, 1–5. https://doi.org/10.1145/3543829.3544519

Paper II

Anna-Maria Meck and Lisa Precht. 2021. How to Design the Perfect Prompt: A Linguistic Approach to Prompt Design in Automotive Voice Assistants – An Exploratory Study. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI ’21). Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/3409118.3475144

Paper III

Anna-Maria Meck. 2023. Secure, Comfortable or Functional: Exploring Domain-Sensitive Prompt Design for In-Car Voice Assistants. In Proceedings of the 5th International Conference on Conversational User Interfaces (CUI ’23). Association for Computing Machinery, New York, NY, USA, Article 45, 1–5. https://doi.org/10.1145/3571884.3604314

Paper IV

Anna-Maria Meck, Christoph Draxler & Thurid Vogt (2023) How May I Interrupt? Linguistic-Driven Design Guidelines for Proactive In-Car Voice Assistants, International Journal of Human–Computer Interaction. 55 pages. https://doi.org/10.1080/10447318.2023.2266251

Paper V

Anna-Maria Meck, Christoph Draxler & Thurid Vogt (2023) Failing with Grace: Exploring the Role of Repair Costs in Conversational Breakdowns with In-Car Voice Assistants, International Journal of Human–Computer Interaction. 71 pages. https://doi.org/10.1080/10447318.2023.2266791

DANKSAGUNG

Dank gilt meinen Betreuer*innen Dr. Draxler, Dr. Vogt und Dr. Precht.
Dr. Draxler für seinen Pragmatismus und seine Hilfestellungen. Und für sein Nachsehen dafür, dass ich den Nominalstil auch durch reichlich Übung nicht losgeworden bin. Dr. Vogt für ihre Ruhe, die Konstanz und die immer klugen Anmerkungen. Und Dr. Precht dafür, alles ins Rollen gebracht zu haben. Vielen Dank an Herrn Prof. Dr. Schmidt für die Übernahme des Zweitgutachtens.

Marion, wenn ich eines Tages nur halb so schnell denken kann wie du, mache ich ein Konvolut an Luftsprüngen. Danke für deine Unterstützung! Der philosophischen Runde den größten Dank für die genaue Kenntnis jeder meiner Gemütslagen und eine unvergessliche Weinreise. Saskia & Maxi, Danke fürs zur Seite stehen. Denise, von Herzen Danke für unsere antizyklische und doch symbiotische Verbindung. Franzi – ich danke dir dafür, immer eine Lösung parat zu haben. Und eine der besten Umarmungen, in deren Genuss ich jemals kommen durfte.
Barbara & Laura. Ihr seid nachhaltig beeindruckende Vorbilder.
Miga & Bilini. Ihr seid der prächtigste Baum, den man sich wünschen kann. Ohne Euch kein Ich.

Omi. Danke für stundenlange Telefonate und deine immerwährenden Ermutigungen.
FK7. Danke für brillantes Entertainment und nimmermüden Techsupport. Who if not you?
Mama & Papa, Ketzl, Butz & Urmeli. Der Metzgamoasta ist für euch.

Last but not least, I wanna thank me
I wanna thank me for believing in me
I wanna thank me for doing all this hard work
I wanna thank me for having no days off
I wanna thank me for never quitting.

Snoop Dog

ABSTRACT

With the rise of Conversational User Interfaces (CUIs), Human-Computer Interaction has made a leap towards natural and intuitive interactions between humans and computers. In the car, CUIs provide a particularly suitable and low-distraction interaction framework. However, inconsiderately designed speech-based interactions with high complexity can result in the opposite effect and increase drivers' cognitive load. CUI designers for in-car interfaces hence face the challenge of developing voice-based interactions for a potentially vulnerable target group in a demanding setting. At the same time, they are not supported by sufficient as well as sufficiently tailored and empirically validated CUI design guidelines. This thesis closes this gap, by answering the research question of how to context-sensitively design CUI prompts and flows in the car. It does so under consideration of various conversational contexts to adequately account for the multi-facetted nature of human interactions. To this end, five research projects develop and validate concrete linguistic-driven design guidelines for CUI prompts and flows.

To determine an efficient way of validating CUI prompts, a first round of studies was conducted to answer the research question of how to efficiently validate in-car prompts in the paper *A Question of Fidelity*. Online crowdsourcing studies emerged as a valid alternative to large-scale driving simulator studies. Subsequently, research identified linguistic parameters with a potential impact on the user experience of CUI prompts in *How to Design the Perfect Prompt*. Three ensuing studies validated the obtained linguistic best practices and prompt design guidelines for different conversational contexts, namely a) the type of interaction, b) the domain of interaction, and c) the initiation of interaction. *How to Design the Perfect Prompt*, *Secure, Comfortable or Functional*, and *How May I Interrupt* showed that CUI prompts need to display a suitable level of (in)formality, complexity/simplicity, and (im)mediacy. Furthermore, an informal, straightforward, and result-oriented speaking style under consideration of the abovementioned contexts is advised. Proactive prompts are thereby specifically dependent on a low level of linguistic complexity as well as a suggestive tone of voice. Additionally, proactive in-car interactions need to carefully consider when to interrupt drivers. To gain insights into best practices for designing CUI flows, the paper *Failing With Grace* explores error handling strategies from Human-Human-Interaction and their applicability to Human-Computer Interaction. The "Principle of Least Collaborative Effort" and concomitant considerations around so-called costs aid CUI practitioners in designing nuanced user-centric and efficient dialog flows for both successful and erroneous conversations.

ZUSAMMENFASSUNG

Mit der Einführung und Verbreitung von Sprachassistenzsystemen hat der Bereich der Mensch-Maschine-Kommunikation einen großen Schritt hin zu natürlichen, größtenteils barrierefreien und intuitiven Interaktionen zwischen Menschen und Computern geleistet. Conversational User Interfaces, kurz CUIs, sind zu einer allgegenwärtigen Technologie geworden, die in verschiedensten Bereichen wie dem Smart Home, dem Gesundheitswesen oder auch in Fahrzeugen verwendet wird. In Fahrzeugen stellen CUIs dabei eine besonders passende Möglichkeit der Interaktion dar. Während die Bedienung von Funktionen via Touch sowohl eine visuelle wie auch eine haptische Ablenkung von der Fahraufgabe bedeutet, bieten Interaktionen via Sprache eine ablenkungsarmere Alternative. Bei der Nutzung von sprachbasierten CUIs können potenziell sicherheitskritische Ablenkungen vermieden werden, ohne jedoch auf die Bedienung von Funktionen verzichten zu müssen [1, 6, 135]. Gleichzeitig können CUIs ihr volles Potenzial nur entfalten, wenn sie spezifisch für diese Modalität gestaltet werden. Anderenfalls, so zeigen Studien, kann der genau gegenteilige Effekt auftreten und Sprachbedienung zu erhöhter kognitiver Beanspruchung führen [132, 135].

Während Sprachassistenzsysteme eine stetig steigende Anzahl unterschiedlicher und vermehrt auch komplexer Funktionen bedienen müssen, werden sie gleichzeitig immer populärer [24]. Interaktionen via Sprache – unserem intuitivsten Weg der Verständigung – führen dabei automatisch zu Vergleichen mit menschlichen Gesprächspartner*innen [34]. Dieser Vergleich führt zu hohen Erwartungen und Anforderungen gegenüber CUIs, den diese aufgrund mangelnder Feedbackmöglichkeiten in Form von beispielsweise visueller Repräsentationen sowie vergleichsweise geringen Kontextwissens häufig nicht erfüllen können [15]. Während CUIs also den Anschein erwecken, natürliche Sprache zu verstehen und mehrstufige kooperative Dialoge führen zu können, bleiben sie oftmals hinter diesem Versprechen zurück [86]. Dabei werden bestimmte Gruppen von Nutzer*innen, wie etwa Dialektsprecher*innen noch immer schlechter verstanden als Sprecher*innen ohne dialektalen Einschlag [8, 54]. Die beschriebene Diskrepanz zwischen den Erwartungen von Nutzer*innen an ein System und dem tatsächlichen Funktionsumfang eines Systems wird als Auswertungs- beziehungsweise Einschätzungsdefizit bezeichnet [103]. Auch dieses Defizit, im englischen „Gulf of Evaluation and Execution“, gilt es im Design von CUIs zu berücksichtigen.

Für ihre Arbeit stehen Designer*innen von CUIs teils zwar in Studien validierte Guidelines für das Design von Konversationslogiken, sogenannten Flows, zur Verfügung, nicht jedoch solche für das Design von Sprachausgaben, sogenannten Prompts. Darüber hinaus sind Guidelines für CUIs häufig an solche für Graphical User Interfaces (GUIs) angelehnt. Im Gegensatz zu CUIs verfolgen GUIs jedoch keinen „Voice First Approach“ und werden nicht per Sprache, sondern via Touch bedient [98]. Weiterhin unterscheiden sich beide Modalitäten hinsichtlich der Möglichkeit, visuellen Input zur Verfügung zu stellen. In CUIs ist es aufgrund verknappter visueller Darstellungsmöglichkeiten beispielsweise wesentlich komplexer, Optionen aufzulisten oder Inputänderungen vorzunehmen [99]. Letztlich gilt es zu beachten, dass Nutzer*innen bei der Bedienung von CUIs ein geringeres Kontrollgefühl erleben als bei der Bedienung von GUIs [83]. Zusammen mit der bereits genannten Entwicklung hin zu höherer technischer Komplexität wird sich dieses Problem weiter verstärken.

Entwickler*innen von CUIs müssen ihre immer komplexer werdenden Produkte also für eine stetig wachsende Anzahl an Nutzer*innen entwickeln, ohne jedoch über empirisch validierte, CUI-spezifische Guidelines zu verfügen. Die vorliegende Arbeit setzt in dieser Forschungslücke an und untersucht und validiert Richtlinien für die Gestaltung von Prompts und Flows in CUIs. Dabei wird ein besonderer Fokus auf Kontextabhängigkeit gelegt, um die komplexe und vielschichtige Natur menschlicher Kommunikation adäquat abbilden zu können. Mithilfe eines linguistisch-empirischen Ansatzes werden konkrete und konkret anwendbare Guidelines für das Design von CUIs validiert. Da das Fahrzeug einen für CUIs besonders passenden Interaktionsraum darstellt, werden Guidelines spezifisch für dieses Setting erarbeitet. Zusammengefasst widmet sich diese Arbeit dem Thema, kontextabhängige Guidelines für das Prompt- und Flowdesign in Sprachassistenzsystemen unter Einbezug der Kontexte a) Gesprächsart, b) Gesprächsinhalt, c) Gesprächsinitiierung und d) Gesprächserfolg zu erarbeiten. Die nachfolgenden Abschnitte beschreiben das gesamthafte Vorgehen der Dissertation und geben einen Überblick über die erhaltenen Ergebnisse.

STUDIENÜBERSICHT

CUI-Designer*innen stehen für die Untersuchung der User Experience gesamthafter CUIs validierte Testmethoden zur Verfügung [57, 67]. Für die Validierung von Systembestandteilen, wie etwa einzelner Prompts, sind solche Methoden jedoch nicht vorhanden. Die erste Studie der

Dissertation widmet sich daher zunächst dem Vergleich unterschiedlicher Testmethoden für die Überprüfung der User Experience von CUI-Prompts.

Studien der User Experience eines Systems sind ein grundlegender Bestandteil von Designprozessen für CUIs [23]. Durch Prototyping können auch frühe und noch nicht gänzlich ausgereifte Produktstände getestet und Produkte basierend auf Feedback von Nutzer*innen angepasst werden. Prototyping kann dabei in Low- und High-Fidelity-Prototyping aufgeteilt werden. Low-Fidelity-Prototypen weisen einen im Vergleich zum Endprodukt geringfügigeren Funktionsumfang auf, wohingegen High-Fidelity-Prototypen diesen bereits annähernd abbilden. Das Erlebbarmachen von CUIs ist aufgrund des iterativen Charakters von Interaktionen mit der hohen Anzahl möglicher Dialogpfade dabei besonders aufwändig. Dieser hohe Aufwand führt zu einer andauernden Debatte hinsichtlich der Sinnigkeit von Low- versus High-Fidelity-Testmethoden [23, 90, 117, 138]. Besonders im Fahrzeugkontext gilt es hier abzuwägen, da High-Fidelity-Methoden neben dem Prototyping eines Interfaces immer auch die Simulation einer konkreten Fahrt und Fahraufgabe verlangen. Während Fahrsimulatorstudien den späteren Use Case unter Ausführung der primären Fahraufgabe zwar genau abbilden, sind sie gleichzeitig kostspielig und zeitintensiv. Eine weniger realitätsnahe, dafür jedoch ressourcenschonendere Alternative kann in Online-Crowdsourcingstudien gefunden werden.

Für das Dissertationsprojekt wurde daher zunächst evaluiert, ob Low-Fidelity-Testmethoden eine valide Alternative zu großangelegten High-Fidelity-Simulatorstudien darstellen. Dafür wurden dieselben 21 Prompts in drei unterschiedlichen Studiensettings in einem Between-Subjects-Design bewertet. Zunächst in Form von zwei Crowdsourcing-Studien, in denen Prompts a) in schriftlicher Form und b) als Audiofile präsentiert wurden und an denen jeweils 75 Proband*innen teilnahmen. Darüber hinaus wurde eine Fahrsimulatorstudie entwickelt, in der 58 Proband*innen Prompts während einer konkreten Fahraufgabe bewerteten. Auf Grundlage der erhaltenen Ergebnisse wurden die weiterführenden inhaltlichen Studien aufgesetzt.

Um das Fundament für die geplanten empirisch-linguistischen Studien zu legen, wurden anschließend syntaktische, grammatikalische und lexikalische Parameter des Deutschen mit einem möglichen Einfluss auf die Bewertung von CUI-Prompts extrahiert [131]. Unterschiedliche Ausprägungen dieser Parameter wurden daraufhin in Online-Crowdsourcingstudien miteinander verglichen. Als erster Kontext wurde dabei zusätzlich die Art des Gesprächs untersucht. Dieser Kontext umfasst unterschiedliche Gesprächsarten, so etwa solche, in denen Nutzer*innen um

Informationen versus die Ausführung einer Funktion bitten. Diesen wurden Gespräche ohne dedizierten aufgabenorientierten Inhalt, sogenannter Chit Chat, gegenübergestellt. In einem A/B-Studiendesign bewerteten 1206 Studienteilnehmer*innen insgesamt 1044 Prompts, um empirisch gesicherte linguistische Best Practices für das Design von CUI-Prompts zu erhalten.

In weiteren Studien wurden die in der ersten Studie erhaltenen Best Practices systematisch für verschiedene Kontexte gegengeprüft. Unterschiedliche thematische Gesprächsschauplätze, sogenannte Domains, bildeten dabei den zweiten Teil der untersuchten kontextabhängigen Beobachtungen. Hier wurden thematisch unterschiedliche Interaktionen – solche mit sicherheitsrelevantem, funktionalem und komfortorientiertem Setting – miteinander verglichen. Grundlage des Studiensettings war ein weiteres Mal ein Between-Subjects A/B-Crowdsourcingformat, in dem 600 Studienteilnehmer*innen insgesamt 162 Prompts bewerteten.

Während Interaktionen mit Sprachassistenzsystemen bislang zumeist vonseiten der Nutzer*innen ausgelöst werden, beinhalten zukünftige Interaktionskonzepte häufig auch proaktive Interaktionen [127, 141]. Auch reaktive versus proaktive Interaktionen wurden als eigenständiger Kontext behandelt und unter dem Gesichtspunkt linguistischer Best Practices beleuchtet. Um einen proaktiven Gesprächskontext adäquat abbilden zu können, wurde eine Simulatorstudie konzipiert. In dieser erlebten 58 Studienteilnehmer*innen 15 proaktive Szenarien in einem Within-Subjects-Design und bewerteten diese anschließend hinsichtlich linguistischer Parameter.

Als letzter Kontext dienten geglückte versus fehlerhafte Interaktionen. Um Guidelines für das Design solcher Interaktionen bereitzustellen, wurde auf Konversationsprinzipien in der Mensch-Mensch-Kommunikation zurückgegriffen. In diesen gilt das sogenannte „Principle of Least Collaborative Effort“, in etwa „Prinzip des geringsten gemeinschaftlichen Aufwands“ [31]. Basierend auf diesem Prinzip streben Kommunikationspartner*innen danach, Fehler in Konversationen mit geringstmöglichem Aufwand zu beheben. In einer Fahrsimulatorstudie erlebten 48 Proband*innen sieben fehlerhafte Dialoge und wurden gebeten, diese qualitativ wie quantitativ zu bewerten. Die Studie wurde als Within-Subjects-Experiment in einem Wizard of Oz-Design konzipiert.

Zusammenfassend ist die vorliegende Doktorarbeit also der Generierung kontextsensitiver Prompt- und Flowdesignguidelines für CUIs im Fahrzeug gewidmet. Diese Guidelines werden im Zuge von fünf Projekten erarbeitet und validiert. Abbildung 2 stellt den Aufbau sowie den Zusammenhang dieser Arbeiten grafisch dar. Da die erhobene Fragestellung die Durchführung großangelegter empirischer Studien erfordert, wurden zunächst effiziente Evaluierungsmöglichkeiten für CUI-Prompts untersucht. Diese Forschung ist in Paper I *A Question of Fidelity. Comparing Different User Testing Methods for Evaluating In-Car Prompts* zu finden. In den nachfolgenden Studien wurden anschließend kontextabhängige Promptdesignguidelines untersucht. Die erste dieser Studien, beschrieben in Paper II *How to Design the Perfect Prompt: A Linguistic Approach to Prompt Design in Automotive Voice Assistants – An Exploratory Study*, ermittelte zunächst linguistische Parameter mit einem Einfluss auf die User Experience von CUI-Prompts in Fahrzeugen. Linguistische Präferenzen wurden dabei kontextsensitiv über verschiedene Gesprächsarten hinweg untersucht. Anschließend wurde ein weiterer Kontext, nämlich der Gesprächsinhalt, in Paper III analysiert: *Secure, Comfortable or Functional: Exploring Domain-Sensitive Prompt Design for In-Car Voice Assistants*. Da proaktive Interaktionen für CUIs immer relevanter werden, stellen sie einen wichtigen Gesprächskontext dar. Daher wurden linguistische Best Practices für das Design proaktiver Prompts im Fahrzeug in Paper IV *How May I Interrupt? Linguistic-Driven Design Guidelines for Proactive In-Car Voice Assistants* näher beleuchtet. Nach der Erarbeitung kontextsensitiver Promptdesignguidelines wurde das letzte Paper dem Design von CUI-Flows gewidmet. In Paper V, *Failing With Grace: Exploring the Role of Repair Costs in Conversational Breakdowns with In-Car Voice Assistants*, wurden Konversationsprinzipien aus der Mensch-Mensch-Kommunikation auf die Mensch-Maschine-Kommunikation angewendet, um Guidelines für die Strukturierung von CUI-Flows zu identifizieren. Die nachfolgenden Absätze geben einen Überblick über die erhaltenen Ergebnisse.

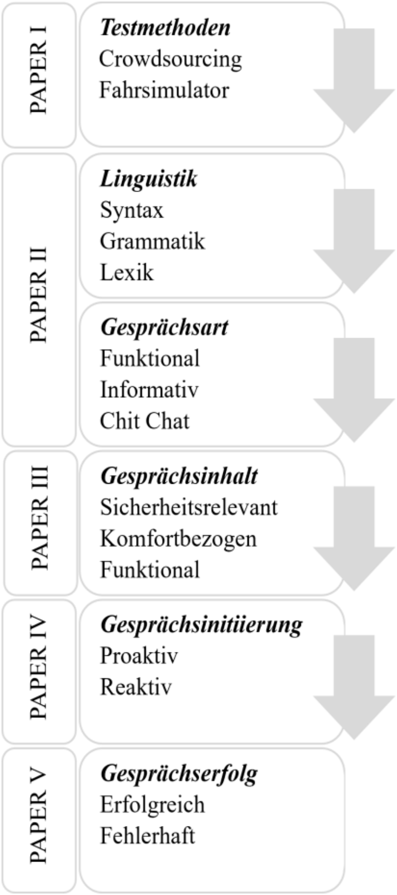


Abbildung 1: Überblick über die in der Dissertation enthaltenen Studien

ERGEBNISÜBERSICHT

Testmethoden
Crowdsourcing
Fahrsimulator

Um natürliche Interaktionen mit CUIs zu ermöglichen, werden Systeme bereits in ihrer Entwicklungsphase getestet. Testmethoden unterscheiden sich dabei stark in Bezug auf die Ressourcen, die für ihre Durchführung benötigt werden. Im Fahrzeugkontext sind aufwändige und teure High-Fidelity-Fahrsimulatorstudien dabei weit verbreitet, gerade für das Abtesten einzelner Prompts jedoch eine umständliche und langwierige Lösung. In einer vergleichenden Between-Subjects-Studie wurde daher untersucht, ob Promptevaluierungen sich zwischen Low- und High-Fidelity-Testmethoden unterscheiden. Dabei konnte kein signifikanter Unterschied in der Evaluierung von Prompts in einer text-basierten Crowdsourcingstudie und einer Fahrsimulatorstudie gefunden werden. Das Ergebnis der Studie wird mit dem Elaboration-Likelihood-Modell erklärt [110]. Das Modell beschreibt unterschiedliche Wege, die von Zuhörer*innen zur Informationsverarbeitung eingeschlagen werden können. Die periphere Route wird dabei gewählt, wenn Zuhörer*innen nicht über ausreichende Fähigkeiten oder die Motivation verfügen, eine Nachricht zu verarbeiten. Auf dieser Route werden statt inhaltlichen Aspekten periphere Hinweisreize in die Verarbeitung einer Nachricht einbezogen. Periphere Anhaltspunkte können als Metadaten einer Nachricht beschrieben werden und beispielsweise die wahrgenommene Glaubwürdigkeit oder Attraktivität einer Quelle umfassen. Die zentrale Route wird dagegen beschritten, wenn Zuhörer*innen ein hohes Wissensbedürfnis aufweisen sowie fähig und motiviert sind, eine Nachricht aufwändig zu verarbeiten [110]. Das erhaltene Ergebnis zeigt, dass Proband*innen in Crowdsourcingstudien wie auch in Fahrsimulatorstudien denselben Verarbeitungsweg einschlagen. Somit stellen Crowdsourcingstudien eine valide und weniger ressourcenintensive Alternative zu Fahrsimulatorstudien dar, wenn es die User Experience einzelner Prompts abzutesten gilt.

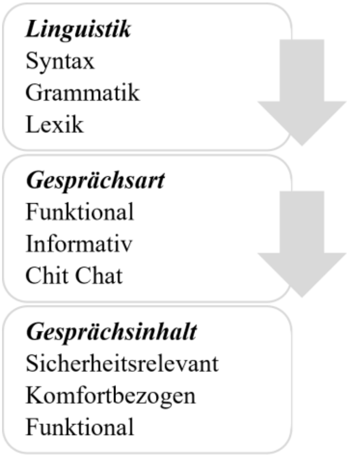
Linguistik
Syntax
Grammatik
Lexik
Gesprächsart
Funktional
Informativ
Chit Chat

Um kontextabhängige Best Practices für die Gestaltung von CUI-Prompts in Fahrzeugen zu erhalten, wurde zunächst eine methodische Untersuchung linguistischer Parameter des Deutschen vorgenommen. Diese ergab 28 syntaktische, grammatikalische sowie lexikalische sprachliche Besonderheiten mit einem potenziellen Einfluss auf die

Bewertung von CUI-Prompts. Diese Parameter wurden in Crowdsourcingstudien miteinander verglichen, um Best Practices für die Gestaltung von Prompts zu erhalten. Gleichzeitig wurden diese Best Practices für unterschiedliche Gesprächsarten, nämlich rein funktionale sowie informative Interaktionen und Chit Chat untersucht. Sowohl auf syntaktischer, wie auch grammatikalischer und lexikalischer Ebene konnten dabei Best Practices abgeleitet werden, die sich teils zwischen Gesprächsarten unterscheiden. Damit ist als gesichert anzusehen, dass die User Experience von CUIs nicht nur von einer passgenauen Dialoglogik, sondern auch von sorgfältig formulierten, kontextsensitiven Prompts abhängig ist. Basierend auf den erhaltenen Best Practices wurden die folgenden CUI-Designguidelines formuliert:

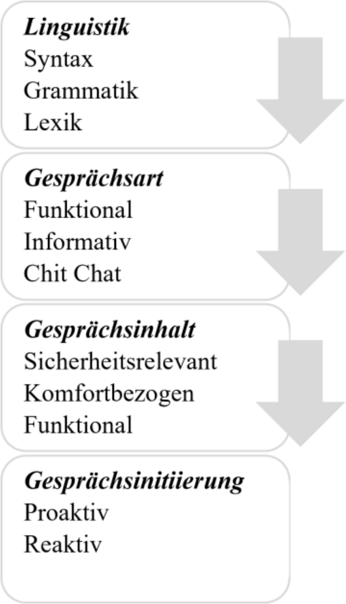
1. Prompts sollten in natürlicher und informeller Sprache geschrieben werden, ohne jedoch zu umgangssprachlich zu sein.
2. Prompts sollten in klarer und einfacher Sprache geschrieben werden und Komplexität vermeiden.
3. Prompts sollten ergebnis- und informationsorientiert geschrieben werden und unnötige sprachliche Bestandteile vermeiden.

Diese Guidelines stellen insofern ein Novum dar, als dass sie methodisch erhoben und maßgeschneidert für den CUI-Kontext validiert wurden. Damit schließen diese Ergebnisse die Lücke der bis dato nur unzureichend vorhandenen Guidelines für CUI-Prompts und bieten ein konkret anwendbares Handbuch für CUI-Designer*innen.

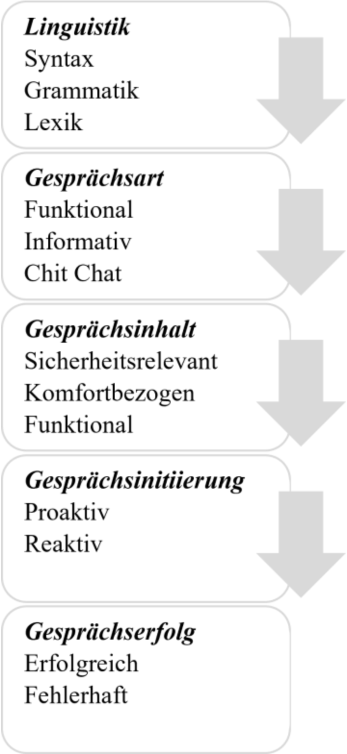


Nachdem die Gesprächsart bereits als relevanter Kontext identifiziert werden konnte, wurde untersucht, inwieweit das thematische Setting einer Konversation ebenfalls einen Einfluss auf linguistische Best Practices aufweist. Als Gesprächsinhalte – sogenannte Domänen – wurden sicherheitsrelevante, rein funktionale sowie komfortorientierte Settings über die obigen Gesprächsarten hinweg miteinander verglichen. Innerhalb Funktionaler Prompts wie auch Informationsprompts konnten dabei keine domänenspezifischen linguistischen Best Practices nachgewiesen werden. Proaktive Interaktionen dahingegen wiesen in Bezug auf einen syntaktischen Parameter domänenabhängige Formulierungspräferenzen auf. Die Ergebnisse dieser Studie zeichnen ein gemischtes Bild hinsichtlich der Bedeutung domänensensiblen Prompt-Designs. Zwar wurden domänenspezifische

Formulierungspräferenzen für proaktive Dialoge nachgewiesen, womit die Bedeutung linguistischer Überlegungen bei der Gestaltung von Prompts unterstrichen wird. Nichtsdestotrotz verlangen Domänen nicht im selben Grad nach nuancierten Formulierungsunterschieden, wie sie für Gesprächsarten nachgewiesen werden konnten. Dieses Ergebnis zeigt, dass linguistische Best Practices abhängig von der Gesprächsart berücksichtigt, Domänen jedoch nur in geringem Maß als weiterer Kontext in CUI-Designguidelines einbezogen werden müssen.



Interaktionen mit Sprachassistenzsystemen entwickeln sich zunehmend von rein reaktiven hin zu proaktiven Systemen. Die Initiierung von Dialogen stellt dabei einen potenziell aufschlussreichen Kontext dar, der in dieser Arbeit in Hinblick auf linguistische Best Practices für das Design von CUI-Prompts genauer beleuchtet werden soll. Während sich bisherige Forschung zu Proaktivität im Fahrzeug auf passende inhaltliche Vorschläge sowie passendes Timing für proaktive „Unterbrechungen" fokussiert hat, wurden konkrete Formulierungen proaktiver Prompts bislang außer Acht gelassen. Wie für die oben bereits beschriebenen Gesprächsarten konnten jedoch auch für proaktive Prompts linguistische Best Practices auf syntaktischer sowie lexikalischer Ebene nachgewiesen werden. Dabei präferierten Studienteilnehmer*innen Prompts mit geringer syntaktischer Komplexität sowie unaufdringlichem Vorschlagscharakter. Dieses Ergebnis unterstreicht, dass das Design-Framework für proaktive Interaktionen um einen linguistischen Kontext erweitert werden muss. Dennoch zeigt die durchgeführte Studie, dass Formulierungspräferenzen für proaktive Prompts weniger ausgeprägt sind als Formulierungspräferenzen für Funktionale Prompts, Informationsprompts und Chit Chat. Die Akzeptanz proaktiver Vorschläge – wiewohl beeinflussbar durch die Beachtung linguistischer Best Practices – hängt zu einem Großteil davon ab, dass Fahrer*innen in passenden Fahrmomenten mit relevanten Vorschlägen angesprochen werden.



Der zuletzt in dieser Arbeit betrachtete Kontext beschäftigt sich mit dem Parameter Gesprächserfolg. Um Guidelines für das Design von CUI-Flows zu erhalten, wurde auf das „Principle of Least Collaborative Effort“ [31] aus der Mensch-Mensch-Kommunikation (HHI) zurückgegriffen. Das Prinzip besagt, dass Kommunikationspartner*innen fehlerhafte Dialoge kostengünstig, also mit geringstmöglichem Aufwand reparieren möchten. Um zu überprüfen, ob das „Principle of Least Collaborative Effort“ auch in der Mensch-Maschine-Kommunikation (HCI) greift, wurden drei dieses Prinzip beachtende HHI-Fehlerhandling-Strategien entwickelt und einer klassischen HCI-Fehlerhandling-Strategie als Baseline gegenübergestellt. In qualitativen Befragungen konnte gezeigt werden, dass Studienteilnehmer*innen Fehlerhandling-Strategien mit niedrigen Reparaturkosten bevorzugen. Kosten stellen somit ein unkompliziertes Messinstrument dar, um die User Experience fehlerhafter Dialoge zu bestimmen. Die erhaltenen Ergebnisse belegen, dass Prinzipien aus der Mensch-Mensch-Interaktion dafür geeignet sind, die User Experience von CUIs zu verbessern und die Mensch-Maschine-Kommunikation natürlicher zu gestalten. Weiterhin führen die Ergebnisse Kosten als Messinstrument und somit Guideline für das Design von CUI-Flows zur Fehlerbehebung ein. CUI-Designer*innen können somit auf einfach ausführbare Kostenberechnungen zurückgreifen, um die User Experience fehlerhafter Dialoge zu bestimmen und diese kostengünstig zu gestalten.

1 INTRODUCTION

With the rise of Conversational User Interfaces (CUIs), Human-Computer Interaction (HCI) has made a leap towards natural, largely barrier-free, and intuitive interactions between humans and computers. CUIs have become a ubiquitous technology with application fields such as the smart home, healthcare, and the car. In the car, CUIs provide a particularly suitable interaction framework for HCI. For most instances, users of in-car CUIs are preoccupied with a demanding primary task: driving. Speech-based interactions allow drivers to keep their hands on the wheel and their eyes on the road. As such, potentially safety-critical distractions can be prevented whilst still allowing control and execution of car functions. In an in-car context, research has found speech-based interactions to be superior to touch-based interactions in terms of efficiency and distractions [1, 6, 74, 75]. With less lane deviation and steadier speed, carrying out interactions via speech leads to a reduction of the cognitive load inflicted on drivers [14]. However, inconsiderately designed speech-based interactions with high complexity result in the opposite and effectively increase drivers' cognitive load [38, 132]. In these cases, speech was found to even "adversely affect traffic safety" [135, p. 1]. Thus, CUI designers for in-car interfaces face the challenge of developing voice-based interactions for a potentially vulnerable target group in a demanding setting.

This challenge will grow in the future, as users foresee their "perfect voice assistants" [141, p. 1] to be smart, personalized, and proactive. Forthcoming CUIs are envisioned to overcome the prevailing "pull paradigm" [127, p. 2] and expected to develop from mostly reactive to progressively proactive assistants with contextual memory. Imminent multimodal features, such as gesture or emotion recognition will thereby prospectively lead to more and more complex interaction patterns. Even today, interactions via language automatically "spark comparisons with human assistants" [34, p. 1]. These comparisons draw on users' experiences from Human-Human Interaction (HHI) and lead to high user expectations towards language-operated interfaces. However, due to current CUIs' "inadequate feedback and impoverished context" [15, p. 19], CUIs are prone to error. While committing and promoting to understand natural language, CUIs frequently fall short of this promise [86]. The discrepancy between limited technical capacities and high user expectations enables the so-called gulf of evaluation and expectations. Coined by Norman [103], the gulf of evaluation and expectation describes the gap between users' expectations of a given system and the actual range of functionalities this particular system offers. The more systems mirror known human-like structures such as language, the more intelligence is thereby expected of them [33, 86]. These high user demands are presently not consistently met by CUIs.

While technical restrictions are currently limiting the full potential of natural interactions with CUIs, CUI designers also face a lack of concrete and CUI-specific design guidelines. Guidelines for CUIs have historically been adapted from guidelines for Graphical User Interfaces (GUIs). While both CUIs and GUIs pose state-of-the-art ways of user interactions with computers and machines, their differences outweigh their similarities. Most GUIs provide users with information in text form, while CUIs are for the largest part operated via speech. Both forms of information processing require the building of a situational model of comprehension [65]. However, compared to text, processing speech places an increased demand on users' working memory [115]. While text can be re-read, speech is fleeting and requires an increased allocation of attention and memory skills [145]. Lastly, GUIs are touch-based and therefore dependent on sequentiality and hierarchies. Speech-based CUIs on the other hand are more exploratory in nature and less structured. As such, merely adapting GUI guidelines to a CUI context means falling short of this interaction medium's high complexity. Already existing guidelines for CUIs partially date back several decades [39, 102, 108, 129]. However, these guidelines have almost exclusively been tailored to defining best practices for designing dialog flows. A dialog flow describes the dialog logic, meaning "the paths that can be taken" [108, p. 23] through a conversation. To allow for a consistent interaction, all possible dialog turns need to be designed as interaction branches to successfully steer a conversation towards completion. The study of CUI prompts on the other hand has received little to no attention. Prompts can be defined as CUI system outputs, meaning the answers a CUI provides. These answers can comprise whole sentences or single word snippets. In case research is specifically concerned with CUI prompt design guidelines, these guidelines are rather general and largely not substantiated empirically. For instance, research and practical experience provided by Pearl [108], Vlahos [140], or Nass & Brave [101] advocates for natural, straightforward, and informal interactions with CUIs. Similar design guidance is provided by Alvarez et al. [1], who argue for "understandable" prompts which should be designed in "short and clear segments" [1, p. 157]. However, the researchers do not define conciseness and clarity further. Moreover, they acknowledge that users' processing capacities are dependent on the current conversational context. Schmidt et al. [124], who center their research around CUIs in an in-car environment, propose precisely and thoroughly formulated system outputs as a best practice when designing prompts. Yet, they do not provide a factual definition of how to comply with their recommendation. Further research producing prompt design guidelines is conducted by Zargham et al. [148]. Their recommendations comprise goal-oriented and concise prompts and are therefore in line with Alverez et al. [1]. Moreover, the researchers describe the optimal prompt as one that is "polite, not imposing, and does not create a feeling of unease" [148, p. 10]. Lastly, Semmens et al. [127] argue for natural interactions between CUIs and users but do not define "natural" further. While the above-mentioned guidelines are sensitive and intuitive, they do not provide factual instructions CUI designers can adhere to. Even fewer studies lend linguistic details for designing CUI prompts. Stier et al. [132, 134], who are pioneering in this area, found users to prefer certain syntactical structures over others though. This research demonstrates the necessity for both linguistic as well as context-sensitive design guidelines.

In sum, designers of CUI experiences are challenged with designing conversations in increasingly complex interaction settings for users with high expectations. At the same time, they are not supported by sufficient as well as sufficiently tailored and empirically validated CUI design guidelines. This thesis attempts to close this gap, by answering the research question of how to context-sensitively design CUI prompts and flows. It does so under consideration of various conversational contexts to adequately account for the complex and multi-facetted nature of human interactions. The structure of the thesis is as follows: chapter 2 provides an overview over aims and objectives of the present work. Subsequently, published papers develop the answers to the proposed research question by examining and validating guidelines for prompt and flow design in CUIs. The thesis closes with conclusions and a summary of specifically tailored CUI design guidelines and practical implications for CUI practitioners.

2 OBJECTIVES AND RESEARCH PLAN

While Conversational User Interface is a collective term for both chat- and voice-based user interfaces, the present thesis is tailored to voice-based interfaces. These interfaces possess the means to process and synthesize natural language. In a first step, a Speech-to-Text (STT) module transcribes spoken language to text. This text is subsequently processed by a component called Natural Language Understanding (NLU). This component parses user intents for key words or phrases and assigns them with meaning. The identified meaning (e.g., starting a navigation in the car) is then assigned to the Dialog Management portion of a system. In this component, the CUI flow as well as CUI prompts are defined to steer a conversation towards successful completion. Lastly, system prompts are synthesized from Text-to-Speech (TTS). Figure 1 shows these functionalities schematically.

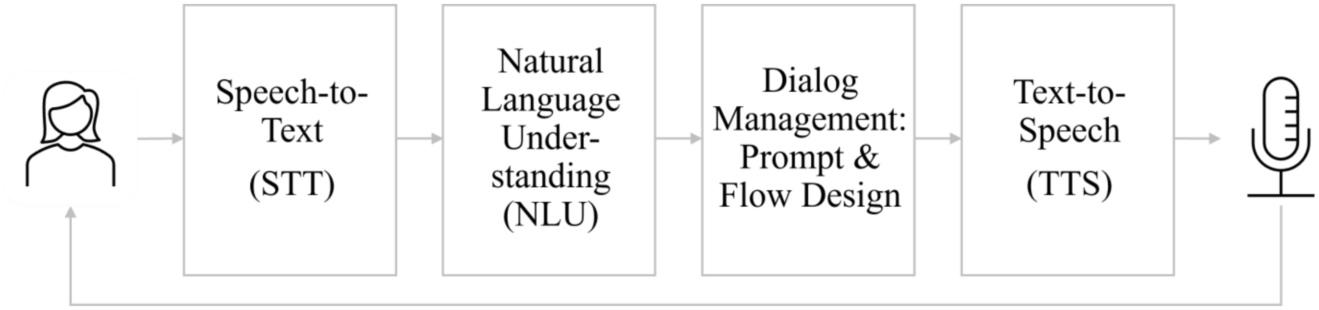

Speech-to-Text (STT)
Natural Language Understanding (NLU)
Dialog Management: Prompt & Flow Design
Text-to-Speech (TTS)

Figure 1: Schematic Functionality of Voice User Interfaces

STT, NLU as well as TTS are technical elements and as such largely not alterable by CUI designers. Still, they do influence the user experience of a given system. For instance, an emotional TTS voice was found to positively impact user emotions and as such measurably impacts CUIs' user experience [61, 100]. The Dialog Management component on the other hand comprises concrete CUI design elements in form of flows and prompts. These design elements can be defined by CUI designers and constitute the most considerable and – more importantly – most employable lever regarding user experience [94, 134]. As such, the present thesis will focus on this Dialog Management dimension to create and validate hands-on best practices and guidelines for the design of CUI prompts and flows.

The already established missing research focus on linguistic-driven prompt design in the field is cause for a lack of tried-and-tested measurement tools for examining the user experience of (in-car) prompts. User studies are a crucial part in user experience design and research, and dependent on validated measurement tools. While a large body of methods, tools and questionnaires exists for evaluating systems as a whole [57, 67, 76], such measures are not available and/or validated on a prompt level. Besides the availability of validated measurement tools, testing methods differ tremendously regarding the timely, organizational, and financial efforts needed to conduct them. User experience research for HCI in the automotive sector is oftentimes conducted in driving simulator studies, which are highly time and cost consuming. While this high-fidelity study format allows the transfer of study results onto real-world driving behavior [143, 144], it is slow and "trades off speed for accuracy" [117, p. 78]. An ongoing debate hence discusses whether less resource-intensive low-fidelity studies can deliver equally valid study results [23, 90, 117, 138].

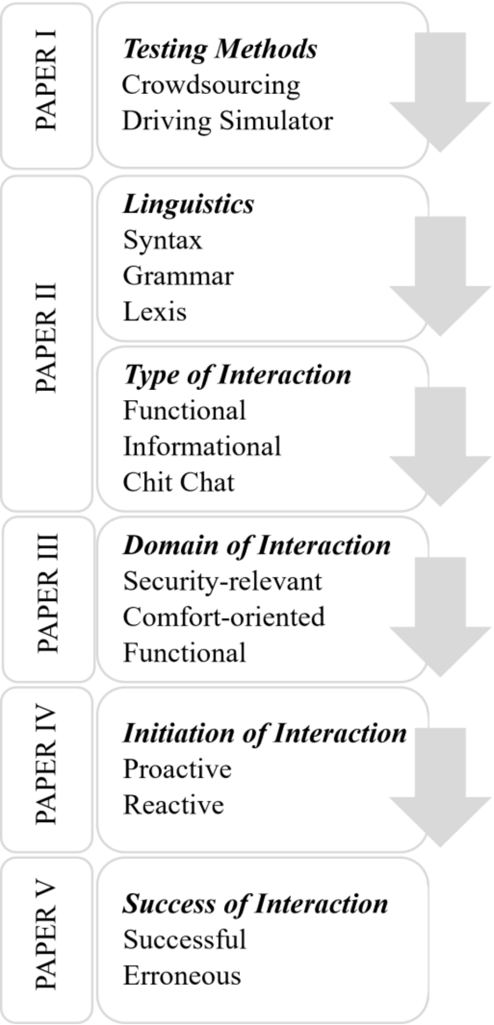
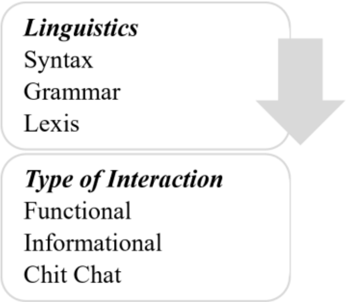


Figure 2: Overview over Studies

This thesis is concerned with developing specifically tailored and concretely applicable guidelines for designing context-sensitive prompts and flows for CUIs in the car. Guidelines are developed and validated by means of five research projects. Together, these projects answer the thesis' research question, namely how CUI prompts and flows can be designed in a context-sensitive manner. Figure 2 displays an overview over the structuring of studies. The proposed research question requires the execution of large-scale empirical user studies, which explore different conversational contexts. Hence, a first round of studies was conducted to answer the research question of how to efficiently validate in-car prompts in Paper I *A Question of Fidelity. Comparing Different User Testing Methods for Evaluating In-Car Prompts*. With an appropriate testing method at hand, content-related studies were executed. The first study, *How to Design the Perfect Prompt: A Linguistic Approach to Prompt Design in Automotive Voice Assistants – An Exploratory Study*, described in Paper II explored linguistic parameters with relevance to the evaluation of in-car prompts. Study participants' linguistic preferences were thereby examined context-sensitively across different types of interactions. Subsequently, a further context, namely the interaction domain, was analyzed in Paper III *Secure, Comfortable or Functional: Exploring Domain-Sensitive Prompt Design for In-Car Voice Assistants*. As proactive interactions are becoming increasingly relevant for CUIs, they pose a most interesting context for in-car interactions. As such, best practices for designing proactive in-car prompts are evaluated in Paper IV *How May I Interrupt? Linguistic-Driven Design Guidelines for Proactive In-Car Voice Assistants*. With the proposition of context-sensitive prompt design guidelines, the last paper was concerned with guidelines for CUI prompts. In Paper V, *Failing With Grace: Exploring the Role* *of Repair Costs in Conversational Breakdowns with In-Car Voice Assistants*, conversational structures from HHI are applied to HCI, delivering guidelines for how to structure interactions with CUIs. The following paragraphs will describe the papers in more detail.

Testing Methods
Crowdsourcing
Driving Simulator

While validated testing methods for assessing the user experience of entire systems exist [17, 57, 67], these methods are currently lacking for single prompts. To close this gap, the first study of this thesis concerned itself with how to efficiently validate in-car prompts. To this extent, various testing methods were compared against each other. Testing methods vary greatly regarding the resources necessary for their execution. Especially in an in-car context, highly resource intensive driving simulator studies are the widespread state-of-the-art testing method. However, considering that prompts represent only a fracture of a CUI, conducting complex simulator studies for the validation of single prompts is a laborious and lengthy approach. Online crowdsourcing studies pose a less resource intensive option to obtain prompt evaluations from a large set of study participants. However, these studies lack the immersive elements of an automotive setting, including the performance of a primary driving task. To see whether online studies pose a valid alternative to driving simulator studies, three comparison studies were set up. First, two crowdsourcing studies with a varying degree of immersiveness were conducted online. These studies were compared to a fully immersive driving simulator study to answer the research question of how to efficiently validate in-car prompts.

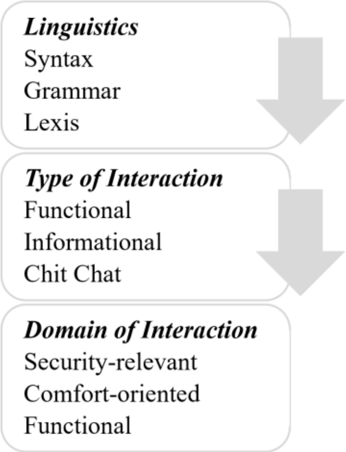


Linguistics
Syntax
Grammar
Lexis

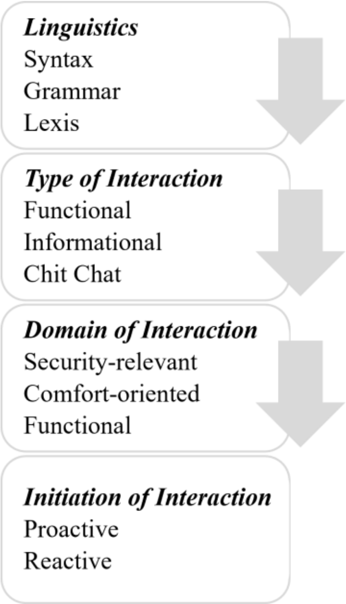
Type of Interaction
Functional
Informational
Chit Chat

After the first study established efficient testing methods for CUI prompts, the second study of this thesis set out to validate CUI prompt design guidelines. Current attempts on such design guidelines have produced intuitive but largely undefined best practices [1, 124, 127, 147]. For instance, research as well as practical experience by experts in the field call for natural, straightforward, and informal interactions with CUIs [101, 108, 140]. However, concrete support on how to adhere to these guidelines is not provided. In contrast, this work focused on developing concretely applicable linguistic best practices on syntactical, grammatical, and lexical levels. To this extent, a German contemporary grammar [131] was searched for linguistic parameters with a potential impact on the evaluation of a prompt. These parameters were subsequently cast into comparison prompts for various types of conversations, namely: a) functional prompts, where users ask a CUI to carry out a function, b) informational prompts, where users ask a CUI for information, and c) chit chat prompts, where users make small talk with a CUI. Comparison prompts were developed for all selected linguistic parameters and conversation types and presented to study participants in an A/B crowdsourcing format. For instance, the study queried whether study participants prefer their CUI to use an active or a passive voice and whether this preference changes depending on the context – in this case the type of interaction. Study participants were presented with two comparison prompts which only differed in one linguistic parameter at a time. In three between-subjects studies, 1206 study participants rated a total of 1044 prompts. The resulting preferred prompt variants constitute operationalized linguistic best practices which can be used to derive empirically validated CUI prompt design guidelines. Throughout the study, the following research questions were answered:

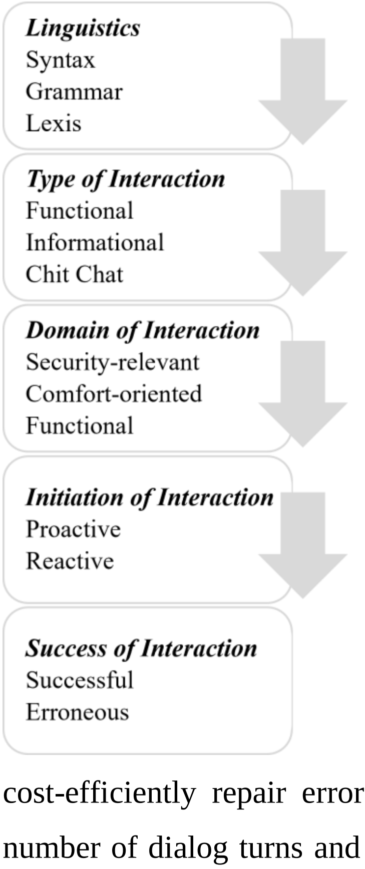
1. What is the entirety of syntactical, grammatical, and lexical parameters with a potential impact on prompt design?
2. Which manifestation of syntactical, grammatical, and lexical parameters is preferred by participants for which prompt type?
3. Which design guidelines and best practices can be distilled on syntactical, grammatical, and lexical levels?
4. Do results allow for identification of overall design patterns?



Interactions with in-car CUIs can span different interaction topics, so-called domains. A considerable portion of interactions is thereby centred around carrying out car functions. However, interactions can also be concerned with e.g., security-relevant, or comfort-oriented content. The previously described studies shed light on the importance and the need for context-sensitive linguistic considerations when designing in-car prompts. Based on these findings, the interaction domain was examined as a further context to develop nuanced linguistic prompt design guidelines. Previous research around interaction domains in an in-car setting showed study participants to prefer different syntactical structures for different domains [134]. However, further research in this field is scarce and an encompassing overview over the importance of domain-sensitive prompt design is still missing. To close this gap, a study was designed to compare linguistic preferences across different interaction domains. For this purpose, comparison prompts were designed for a) security-relevant, b) functional, and c) comfort-oriented interactions and compared in an A/B study. As the type of interaction proved to be a considerable context, linguistic preferences were observed for functional prompts, informational prompts, and proactive prompts. Proactive prompts replaced chit chat prompts in this study, as interaction domains are not applicable for small talk. Comparison prompts were varied systematically regarding syntactical, grammatical, as well as lexical parameters. In a between-subjects A/B study format, 600 study participants compared 54 linguistically altered prompts across interaction types and interaction domains. Throughout the study, the following research question is answered: Which effect do domains have on the preference for linguistic parameters across dialog types?



With CUIs being expected to develop from mostly reactive to proactive assistants [95, 104, 114, 141], the initiation of interactions becomes an intriguing conversational context. Designing successful proactive interactions has thereby proven to be dependent on a multitude of factors. In interactions with reactive agents, users can define content and timing of conversations themselves. Proactive agents on the other hand potentially disturb users who are engaged in primary tasks. In the automotive context, this primary task is likely security-relevant and makes precise timing of proactive interactions of paramount importance. While related research is already concerned with *when* to proactively address users and drivers [72, 114, 122, 127], the linguistic dimension of *how* to design proactive prompts has not been touched upon yet. The previous studies have shown the importance of context-sensitive prompt design though. Prompts can produce varying degrees of language complexity, crucial for the management of drivers' cognitive load. Furthermore, proactive interactions are to be suggestive rather than imposing [114, 148]. This requirement can potentially be controlled by an appropriate tone of voice in form of linguistic considerations. To examine best practices for designing proactive in-car prompts, a within-subjects driving simulator study was conducted. The change of study medium from crowdsourcing study to driving simulator study was made due to findings from related research around in-car proactivity. Previous studies have found a strong link between primary task engagement and the acceptance of proactive interactions. To respect this correlation, a driving simulator study was designed. This change of study environment required the revision of previously used research methods as well as the streamlining of study parameters. As A/B studies are not a feasible method for comparing prompts while driving, a Likert scale for the evaluation of CUI prompts was developed and validated. This scale enhances the currently limited pool of validated evaluation methods for single prompts. While the previously conducted crowdsourcing studies allowed for the examination of a large number of prompts, parameters were consolidated for the simulator study. Minor and highly specific parameters as well as parameters with very clear preferences were no longer evaluated. In the driving simulator study, 58 study participants experienced 15 proactive use cases with varying linguistic formulations to answer the following research question: are there best practices for the design of proactive prompts?



The papers described above conclude the thesis' work on prompt design guidelines. Linguistic and context-sensitive considerations emerged as an integral building block for the user experience of in-car CUIs. The last project of the thesis focused on CUI flow design guidelines. While technological advancements continue to improve CUIs, interactions with assistants are still prone to error. Some user groups, such as dialect and vernacular speakers are affected disproportionally by this problem [8, 54]. Errors are thereby not confined to HCI but are a constant in HHI too. To examine a last conversational context and derive guidelines for the design of CUI flows, this thesis analyzes erroneous conversations. It does so by drawing from error handling strategies which are commonly applied in conversational breakdowns between humans [31]. Here, human interlocutors apply the "Principle of Least Collaborative Effort" to cost-efficiently repair errors, such as misunderstandings. Costs are thereby composed of the number of dialog turns and words necessary to repair a dialog. To examine to which extent the Principle of Least Collaborative Effort and cost calculations must be afforded a place in HCI, a within-subjects driving simulator study was conducted. In this study, 48 study participants experienced seven erroneous conversations and were asked to rate error handling strategies quantitatively as well as qualitatively. The study answered two research questions, namely whether different error handling strategies vary regarding their repair costs and whether the costs associated with repairing errors influence the preference for error handling strategies.

3 PAPER

I. A QUESTION OF FIDELITY. COMPARING DIFFERENT USER TESTING METHODS FOR EVALUATING IN-CAR PROMPTS[1]

User studies are a major component in any user-centered design process. Testing methods thereby vary tremendously regarding the organizational, financial, and timely effort needed to conduct them. Driving simulator studies generally are the method of choice when dialogs need to be validated for in-car settings. These studies are highly time- and cost-consuming though. Online crowdsourcing studies can be an alternative as they allow for quick results and large sample sizes while at the same time being time- and cost-efficient. Still, voice user interface designers argue for a lack of applicability to concrete use cases. This is especially true for speech dialog systems in an in-car context where users experience voice as a secondary task with the primary task being driving. To compare the validity of different user testing methods, study participants in a between-subjects study design evaluated proactive in-car prompts presented a) in an online crowdsourcing study in text form, b) in an online crowdsourcing study via audio, and c) in a driving simulator. Prompt evaluations did not differ significantly between conditions a) and c) but diverged for condition b). Findings are explained by drawing from the Elaboration Likelihood Model and used to answer the question of how to efficiently validate in-car prompts.

1 INTRODUCTION

Testing is a crucial part of any design journey and as such stretches into the realm of designing HCIs. Testing methods like the elicitation model and Wizard of Oz studies allow for extracting potential interaction paths a given user will later adopt to navigate a system [23]. Besides testing the dialog flow, concrete voice assistant system outputs (prompts) need to be evaluated to determine whether a voice user interface (VUI) is communicating intuitively, clearly, and efficiently with users – even more so as studies have shown that users prefer certain prompt formulations on syntactical, grammatical, and lexical levels [94, 134]. This is especially important for in-car use cases as conversations may not distract drivers from their primary driving task [135]. At the same time, the iterative nature of conversations with a large number of possible dialog paths

[1] As the leading author, I developed the research idea as well as the experiment design and conducted and analyzed all described studies. Dr. Christoph Draxler and Dr. Thurid Vogt supported the development of the research idea and the experiment design and provided feedback on the overall work.

makes prototyping for voice an effortful task [68]. This high effort leads to an ongoing and vivid debate as to whether low-fidelity prototypes (meaning prototypes with limited functionality and interaction possibilities, not accurately representing a later use case) pose a valid option to conduct user testings and obtain meaningful results when testing VUIs [23, 90, 117, 138]. Where low-fidelity prototypes lack the extent and interaction possibilities of a later system, high-fidelity prototypes provide just that: users can interact with them as with the real product. Still, they "trade off speed for accuracy" [117, p. 78] and are highly time-, cost-, and personnel-intensive. This is particularly true for an automotive context where a high-fidelity prototype means simulating a concrete driving situation and providing a physical car mock-up on top of prototyping a (voice) user interface. While a driving simulator study certainly provides the most accurate setting for conducting user studies to evaluate in-car use cases, it is also highly strenuous regarding resources. Low-fidelity alternatives include less time- and cost-intensive online crowdsourcing studies. Taking a driving simulator study as the baseline, we conducted two online crowdsourcing studies, one in written form and one auditorily via sound files, to examine whether prompt evaluations differ significantly between these conditions. The aim of the present study is to determine whether low-cost and easily set up low-fidelity testing methods constitute a valid alternative for high-fidelity testing of voice prompts for in-car contexts.

2 DRIVING SIMULATOR VS CROWDSOURCING STUDIES

A driving simulator study provides a high-fidelity framework for conducting user studies, albeit at the expense of high financial and organizational efforts. The following paragraphs are to compare driving simulator studies and online crowdsourcing studies based on four parameters: (low- vs high-fidelity) environment, evaluation task, working memory demand, and carrier medium.

Study environments differ regarding their fidelity, meaning how well they manage to replicate a later product. Cambre and Kulkarni give a profound overview over prototyping methods and tools for VUIs with different levels of fidelity [23]. Low-fidelity methods include elicitation methods to gather insights into how users will later use a product. High-fidelity environments include Wizard of Oz studies or functional NLU platforms with the ability to process and understand speech. For in-car speech contexts, a driving simulator equipped with a voice assistant is the most high-fidelity testing environment. Online crowdsourcing studies on the other hand lack the car context and do not elucidate actual first-hand interactions with a VUI. Simulator and crowdsourcing studies also differ regarding the types of tasks they impose on study participants. Study participants in driving simulator studies evaluate dialogs or prompts as secondary tasks while their primary task is driving. This creates a dual-task environment, being known for demanding additional cognitive load [42]. Such a primary task is absent in crowdsourcing studies, where handling concurrent tasks is not necessary. Study participants in different testing conditions thus experience different working memory demands.

On top of the presence or absence of a primary task, working memory demands differ regarding the presentation of prompts. Participants in high-fidelity simulator conditions listen to prompts while participants in crowdsourcing conditions oftentimes only read them. Processing of spoken and written language is handled differently from working memory side though which poses a considerable difference between the testing methods. While readers are able to comprehend 238 words per minute [19], listeners are most comfortable with processing spoken language at a rate of 180 words per minute [115]. Both reading and listening require resources to build a mental model of the read/heard contents, the so-called situation model of comprehension [65]. While text can be re-read, speech is fleeting, therefore requiring additional attentional and memory skills [145].

Lastly, the three conditions differ regarding the carrier medium, meaning the channel delivering prompts to study participants. Prompts can be conveyed via audio or via text which results in the presence respectively absence of a text-to-speech (TTS) voice. Studies have shown that synthetic voices influence perceptions of virtual assistants in regard to e.g. building trust and behavioral intentions [28]. Furthermore, synthetic voices are found to be less understandable and appealing than human voices [20]. Hence, an influence of TTS on evaluation of prompts is highly likely. The following table shows the three testing conditions driving simulator, crowdsourcing audio, and crowdsourcing text as well as the four parameters they differ in. Green marked cells show parameters in which conditions overlap.

Table 1: Differences and Similarities between Testing Conditions

	Driving Simulator	**Crowdsourcing Audio**	**Crowdsourcing Text**
Environment	High-fidelity	Low-fidelity	Low-fidelity
Evaluation Task	Secondary task	Primary task	Primary task
Working Memory Demand[2]	High (Listening Task)	High (Listening Task)	Low (Reading Task)
Carrier Medium	Speech/TTS	Speech/TTS	Text

[2] Please note that “High” and “Low” are not meant in absolute but relative terms to highlight differences in working memory demand between testing conditions.

Originating from this assessment and conditions' similarities and differences, we formulated the following hypotheses for our study:

H1: Users' prompt evaluations differ significantly between the text and the driving simulator condition.

H2: Users' prompt evaluations do not differ significantly between the text and the audio condition.

H3: Users' prompt evaluations do not differ significantly between the audio and the driving simulator condition.

3 METHOD

To compare the conditions crowdsourcing text, crowdsourcing audio, and driving simulator, a between-subjects study design was set up. The study format was selected to minimize carry-over, learning and habituation effects [48]. In total, 21 already existing proactive prompts of BMW's Intelligent Personal Assistant were used as study prompts and presented to study participants a) in an online crowdsourcing study in text form, b) in an online crowdsourcing study via audio, and c) in a driving simulator. A proactive use case was chosen to create a setting with a one-shot prompt to ensure controllability of dialog turns and sequences.

All participants completed a set of demographic questions before moving on to the evaluation task. Across conditions, participants were asked to rate the prompts they heard respectively read on four seven-level Likert scales: 1. very positive – very negative, 2. very intelligent – very naïve, 3. very simple – very complicated, 4. very natural – very unnatural. In order for study participants to focus on the contents and formulations rather than the synthetic voice speaking a prompt, study participants in the simulator and the audio condition were encouraged to ignore the TTS in their ratings as good as possible [134]. Prompts were presented to study participants one prompt at a time and in randomized order to counteract sequence effects. Study participants in the audio and the simulator condition heard every prompt once without the possibility to revisit a previous prompt. BMW's current female synthetic voice served as TTS output for the study and was used for both the simulator and the audio condition. Evaluations in all testing conditions were collected directly after study participants heard/read a prompt. The study was conducted in German.

3.1 Online Crowdsourcing Studies

Two studies – one text-based and one audio-based study – were designed and distributed online via crowdsourcing [36]. Both studies were deliberately not equipped with elements to immerse study participants further into a driving scenario to uncloudedly compare low- and high-fidelity settings. Study participants received an introductory text asking them to imagine driving a vehicle accommodated with a proactive voice assistant. Study participants for both text- and audio-based studies were asked to imagine experiencing the prompts they read respectively heard in an in-car setting. For each prompt, a short introductory text was provided to explain the context in which a given prompt would occur in an in-car scenario, e.g. *"Your voice assistant proactively points out the 'Parking' function. Please rate the prompt on the scales."*

3.2 Simulator Study

The driving simulator study was conducted in a simulator with a 180° screen and a stationary vehicle mock-up. The mock-up was equipped with a voice assistant managed by the experimenter in a Wizard of Oz setting. A highway setting with low traffic was selected as driving scenario to limit cognitive load from the primary driving task. This approach stems from findings by Stier and Sigloch, indicating that a too complex driving task can overshadow prompt evaluations [134]. Each participant conducted a five-minute familiarization drive to get to know the vehicle and the route. Afterwards, they were introduced to the task. Participants were asked to follow a lead vehicle with 100km/h on the right lane and rate proactive voice prompts directly after hearing them.

3.3 Participants

A total of 208 native German speaking participants completed the studies with 75 participants each for the text and the audio condition as well as 58 participants for the driving simulator condition. Two participants in the simulator condition terminated the study themselves due to motion sickness. To check for homogeneity of the samples, Pearson's Chi-Squared Tests for Homogeneity were conducted regarding study participants' age, sex, experience with in-car voice assistants, and the number of kilometers travelled by car per year. Participant groups in all conditions proved to be homogeneous with p-values ranging from a minimum of p=.1 to a maximum of p=.77.

4 RESULTS

A post-hoc power analysis was conducted with G*Power [43] to determine if the sample size was sufficient which proved to be the case: α=.05, sample size: 208, effect size f: 0.29 → β=0.98. To visualize results, the following boxplot was drawn. Overall, study participants rated prompts in the audio condition more poorly (mean=3.38, sd=1.6) than in the simulator (mean=2.6, sd=1.3) and the text condition (mean=2.5, sd=1.31)[3]. A Cumulative Link Mixed Model using the ordinal package [30] was then fitted to account for the ordinal structure of the data in R [116]. Study participants' evaluations (dependent variable) were predicted as a function of the respective testing condition. The between-factor *condition* (audio vs text vs simulator) therefore served as fixed factor while within-factors *participants* and *prompts* (rated system outputs) were introduced as random factors with random intercepts and random slopes. A model with random slopes and random intercepts was chosen because of the repeated-measures character of the variables *participants* and *prompts*. A follow-up AIC model selection confirmed this choice with the model with random intercepts and random slopes carrying 100% of the cumulative model weight compared to a model without random intercepts and random slopes. The full model translates to *clmm(evaluation ~ condition + (1+ condition|participants) + (1+condition|prompts))*.

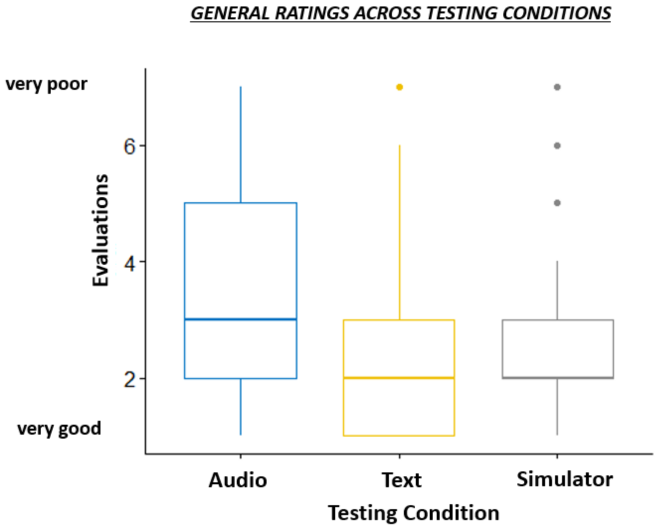


Figure 1: General Ratings Across Testing Conditions

Results are reported for α=.05. The model's total explanatory power is substantial (conditional R^2 = 0.338) with marginal R^2 = 0.083. Within this model, *condition* proved to be significant: OR 0.28, 95% CI [0.19-0.42], p<0.001. Post-hoc Tukey's HSD Tests for multiple comparisons were conducted to see which groups were significantly different from each other. Evaluations of participants differed significantly between the simulator and the audio condition (z.ratio 5.644 , p <.0001) as well as the text and the audio condition (z.ratio 7.381, p <.0001). There

[3] Please see '5 Discussion' for the discussion of Naturalness ratings.

were however no statistically significant differences in study participants' evaluations between the text and the simulator condition (z.ratio -0.807, p=.7). Hypotheses 1-3 can therefore not be supported and need to be rejected.

5 DISCUSSION

In this study, we compared different testing conditions in high- vs low-fidelity surroundings. We found that there is a significant impact of the testing condition on users' prompt evaluations, albeit different from what was expected. Unlike formulated in the hypotheses, prompt evaluations did not differ for the *text-simulator* condition but were significantly different for the *audio-text* as well as the *audio-simulator* condition. These findings – while surprising at first – can be accounted for by drawing from research around the Elaboration Likelihood Model. Introduced by Petty and Cacioppo, the Elaboration Likelihood Model (ELM) explains "communication-induced attitudinal change" [110, p. 125] by defining different so-called routes a given message recipient can take: a) the central route and b) the peripheral route of processing. This dichotomy results from a person's ability and motivation to process a given message and can also be described as controlled (central route) and automatic (peripheral route) processing. Central processing is based on the "true merits of the information presented" [110, p. 125] while peripheral processing leans on so-called affectional or peripheral cues of a message which are less content- and more context-based. Peripheral cues can be described as the metadata surrounding a message, e.g., perceived credibility or attractiveness of a source or – in the case of this study – a TTS voice. The ability and the motivation to process a message are central elements to the ELM. If ability as well as motivation are given, recipients are willing to allocate "considerable cognitive resources" [110, p. 128] to process a message and will take the central route of processing. If recipients are not motivated and/or lack the ability to process a message, they will exit the central route and enter the peripheral route of processing instead. In doing so, they will rely more heavily on peripheral cues of a message when processing it.

Our results suggest that study participants in the simulator and the text condition used the central route of processing while participants in the audio condition processed prompts by taking the peripheral route of processing. The general ability to process prompts was given in all three testing conditions. All participants rated the same 21 prompts on four seven-level Likert scales. The motivation as well as the possibility to fall back on peripheral cues is the differentiating element between the testing conditions. The concrete, high-fidelity driving environment in the simulator condition increased the personal relevance for study participants to carefully evaluate prompts and provided them with the motivation to allocate considerable cognitive resources to this task. Thus, these participants took the central route of processing.

With the omission of the primary driving task in the low-fidelity online crowdsourcing studies, the imminent relevance of prompts as well as the motivation for the more cognitively demanding central processing were withdrawn. Study participants in the audio and the text condition were now susceptible to taking the peripheral route of processing. The audio and the text condition differ in two key points though: the amount of mental workload (= the considerable cognitive resources) needed for processing a prompt and the presence respectively absence of peripheral cues. Despite reduced motivation compared to the simulator condition, reduced attentional and memory skills needed to process text compared to speech [145] fostered taking the central route of processing for study participants in the text condition. In contrast, reduced motivation, the higher need for using considerable cognitive resources, and the presence of peripheral cues in form of a TTS voice led study participants in the audio condition down the peripheral route of processing. These study participants were hence more susceptible to peripheral cues, which is substantiated when comparing the evaluations of naturalness across testing conditions. Study participants for the audio and the simulator condition were specifically asked to ignore peripheral cues in form of the TTS voice and focus on contents and formulations when evaluating prompts. Our results suggest that participants in the simulator condition were indeed able to ignore peripheral cues because evaluations for naturalness do not differ significantly between the *simulator-text* condition (z.ratio 0.373, p=0.9262). Study participants in the audio condition on the other hand rated prompts significantly less natural (*audio-simulator*: z.ratio 10.390, p<.0001, *audio-text*: z.ratio 10.496, p<.0001), suggesting that they took peripheral cues into account when evaluating prompts. This can be

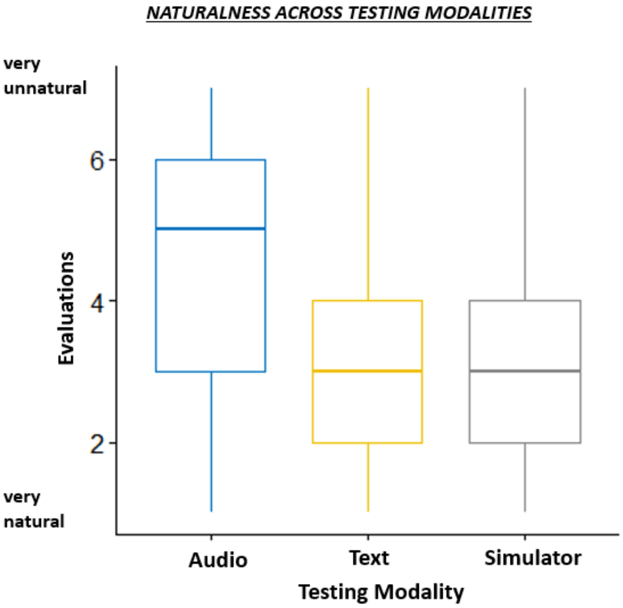


Figure 2: Naturalness Ratings Across Testing Conditions

seen in Figure 2 where naturalness ratings for the audio condition are significantly poorer compared to naturalness ratings for the simulator condition (z.ratio 10.390, p <.0001). This indicates that study participants in the audio condition not only rated the prompt but also peripheral cues, namely the TTS voice speaking the prompts.

Summing up this paper's findings: prompt evaluations do not differ significantly between a simulator and an online crowdsourcing study conducted in text form, because study participants process prompts similarly in these conditions. The following figure displays the two processing possibilities study participants have at their disposal and shows which route study participants in all three testing conditions took when evaluating proactive in-car prompts.

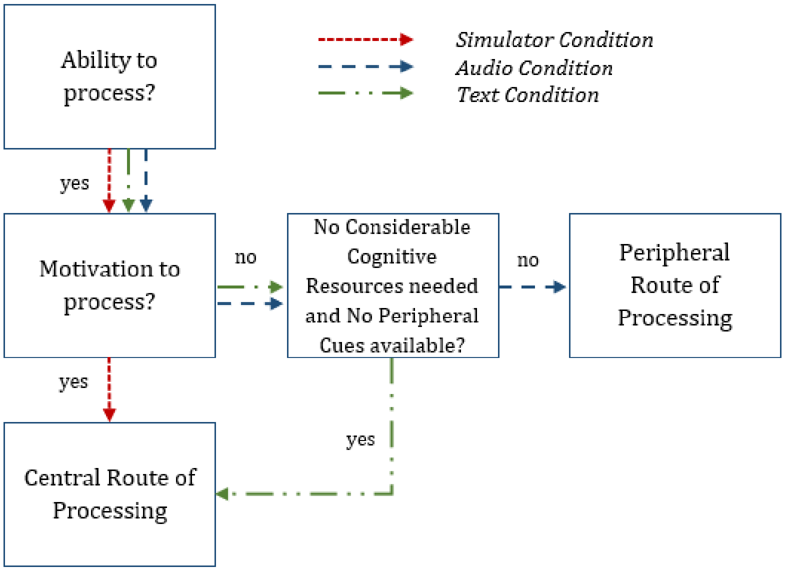

Ability to process?
Simulator Condition
Audio Condition
Text Condition
yes
Motivation to process?
no
No Considerable Cognitive Resources needed and No Peripheral Cues available?
no
Peripheral Route of Processing
yes
yes
Central Route of Processing

Figure 3: Processing Routes Across Testing Conditions

6 LIMITATIONS & FUTURE RESEARCH

The present study did not examine prompt evaluations in multi-step dialog contexts but for proactive one-shot prompts. While results showed no significant differences in prompt evaluations between a driving simulator study and a corresponding crowdsourcing study conducted via text, the same does not necessarily have to be true for prompts in multi-step dialogs. For multi-step scenarios, there is a large number of further factors like e.g., NLU mismatches, latencies, etc. with a potential impact on the evaluation of a prompt. These cannot as easily be replicated in crowdsourcing studies which makes comparability with high-fidelity simulator studies potentially more difficult. Future research can build on this paper's results by enhancing and refining testing methods in crowdsourcing environments though. “High-fidelity” elements such as videos, and/or a primary task could immerse study participants further into an in-car environment, thereby better portraying the concrete later use case. This paper’s results were explained by drawing from the ELM, which – to the best of our knowledge – has not yet been applied to similar research. Future work could explore this further by carrying out studies specifically aimed at incorporating the model.

7 CONCLUSION

There is a considerable need for UX testing in the field of voice user interface design to deliver on the promise of natural interactions. Testing methods thereby vary substantially regarding the financial and organizational efforts needed to conduct them. While high-fidelity testing environments are especially popular for validating in-car use cases, this study showed that prompt evaluations did not differ significantly between a simulator study and an online crowdsourcing study where prompts were presented in text form. Crowdsourcing studies thus constitute a valid and less time- and finance-consuming alternative for simulator studies if the suitability of single prompts needs to be confirmed through user testing.

II. HOW TO DESIGN THE PERFECT PROMPT: A LINGUISTIC APPROACH TO PROMPT DESIGN IN AUTOMOTIVE VOICE ASSISTANTS – AN EXPLORATORY STUDY[4]

A Linguistic Approach to Prompt Design in Automotive VAs for German

In-vehicle voice user interfaces (VUIs) are becoming increasingly popular while needing to handle more and more complex functions. While many guidelines exist in terms of dialog design, a methodical and encompassing approach to prompt design is absent in the scientific landscape. The present work closes this gap by providing such an approach in form of linguistic-centered research. By extracting syntactical, lexical, and grammatical parameters from a German contemporary grammar, we examine how their respective manifestations affect users' perception of a given system output across different prompt types. Through exploratory studies with a total of 1,206 participants, we provide concrete best practices to optimize and refine the design of VUI prompts. Based on these best practices, three superordinate user needs regarding prompt design can be identified: a) a suitable level of (in)formality, b) a suitable level of complexity/simplicity, and c) a suitable level of (im)mediacy.

1 INTRODUCTION

The in-car environment provides the optimal framework for speech control. The possibility for drivers to keep their hands on the wheel and their eyes on the road makes maneuvering functions with a Voice Assistant (VA) more efficient, less error-prone, and less distracting than carrying them out manually. Studies find less lane deviation and steadier speed for participants executing functions via voice when compared to touch [14]. Furthermore, this operating mode reduces drivers' cognitive load, not distracting them from their primary driving task [1, 6, 135]. Designed inconsiderately though, a reverse effect can be observed. Studies show an increase in cognitive load when prompts (i.e., VA system outputs, e.g., in form of "Okay, I'll start the navigation right away. Your next destination is Munich") are designed too complexly, e.g. in terms of an intricate syntactical structure [38, 132]. The same effect can be observed when applying voice for improper use cases involving high cognitive demand. Studies suggest that VA usage can even 'adversely

[4] As the leading author, I developed the research idea as well as the experiment design and conducted and analyzed all described studies. Dr. Lisa Precht supported the development of the research idea and the experiment design substantially and provided feedback on the overall work.

affect traffic safety’ in these situations [135, p. 1]. As of today, in-car voice user interfaces (VUIs) are oftentimes designed based on GUI solutions not pursuing a voice first approach [98]. This adds to the above-mentioned issue as both interfaces differ in many regards. Visual aids in VUIs are reduced or entirely absent in comparison with GUIs, making it harder to convey information. Additionally, the lack of a visible hierarchical structure makes revisions and edits more difficult for users [99]. Regarding the operation mode, a diminished sense of agency for users interacting with a speech interface compared to a keyboard interface can be found. A study by Limerick et al. links this to the increased cognitive working memory load accompanying the use of speech [83]. With technical advancements supporting more and more complex use cases via voice in the future, this problem will intensify. In addition, the number of users of in-car VAs is still on the rise, increasing the amount of in-car conversations overall [24]. Designers of in-car voice experiences are thus facing the problem of designing for a surging number of users whilst handling the requirements of increased technical complexity without sufficient guidelines.

Research has been conducted as to how a conversational user interface needs to be designed in terms of best practices for dialog guidance and dialog management, covering structural as well as technical aspects of voice design [13, 74, 75, 108, 135]. The design of prompts on a linguistic-centric level has received less attention. To the extent of the authors’ knowledge, the composition of system outputs has not been studied on a broad linguistic spectrum. Language-dependent syntactical, grammatical, and lexical parameters influence drivers and their driving performance though [133]. It is therefore crucial that the design of VA system outputs is being carried out attentively. Moreover, Stier et al. exposed different user preferences for syntactical structures when comparing them across in-vehicle use cases [132]. In addition to these use case effects, the type of prompt also has a potential impact on its evaluation. In order to investigate this, we propose a three-part cluster for in-vehicle prompts: 1) Functional Prompts: confirming function execution and asking for user input, 2) Informational Prompts: informing users about in-vehicle functions, and 3) Chit Chat Prompts: chatting informally and independently of function.

When designing voice experiences, already established design principles need to be enhanced with linguistic parameters as well as considerations around the prompt type. This paper presents results of exploratory studies focusing on these linguistic parameters and their influence on prompt evaluation. The study format was chosen due to a lack of previous research in this field. The paper aims to close this gap by extracting syntactical, lexical, and grammatical parameters and examining how their respective manifestations affect study participants' perception of a given system output across different prompt types. Based on this thorough linguistic approach and the derivation of best practices in prompt design, this work intends to develop operationalizable guidelines for the design of VUIs by answering the following research questions:

Research Question I: What is the entirety of syntactical, grammatical, and lexical parameters with a potential impact on prompt design?

Research Question II: Which manifestation of syntactical, grammatical, and lexical parameters is preferred by participants for which prompt type?

Research Question III: Which design guidelines and best practices can be distilled on syntactical, grammatical, and lexical levels?

Research Question IV: Do results allow for identification of overall design patterns?

2 METHOD

2.1 Extracting Study Parameters

To obtain syntactical, grammatical, and lexical parameters with a possible influence on prompt evaluation, a complete German contemporary grammar was analyzed [131]. All the grammar's 94 chapters with 34 chapters concerned with syntax, 28 chapters dealing with grammar, and 32 chapters dealing with lexis were studied. The chapters were each concerned with concrete syntactical, grammatical, and lexical parameters of German. These parameters were extracted and examined regarding their ability to form distinguishable, yet comparable manifestations. To give an example: the chapter 'Voice' (German 'Genus verbi') is concerned with the parameter 'voice' which has two manifestations, namely active voice and passive voice.

A high number of parameters thereby proved to be insufficient for the studies' purpose and was hence deleted. Reasons for deletion were a) inability of a parameter to form two or more comparable manifestations, b) alternation in manifestation leading to an alternation in meaning, c) alternation in manifestation affecting further syntactical, grammatical, or lexical parameters (therefore making comparisons ambiguous), and d) inability of a parameter to be expressed across the prompt types Functional Prompts, Informational Prompts, and Chit Chat Prompts. After filtering all extracted parameters regarding these criteria, a total of 28 parameters met the requirements to be tested in studies. Subsequently, these were cast into study prompts.

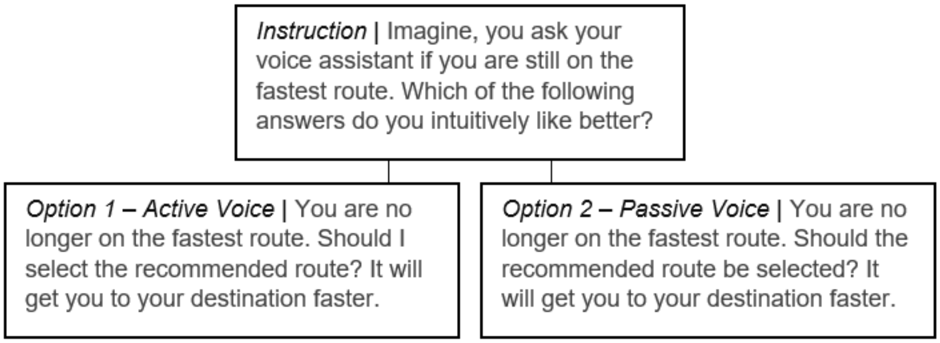
2.2 Designing Study Prompts

To expose which manifestation (e.g., active voice vs passive voice) of a given parameter (e.g., voice) is preferred by study participants, two comparison prompts were designed for each of the extracted parameters. Already existing system outputs of BMW's Intelligent Personal Assistant served as basis for stimuli for the present study to ensure consistency of the VA's persona and to display actual in-vehicle use cases. For each parameter, a minimum of three pairs of exemplary prompts was designed to serve the exploratory nature of the study and to find a broad range of possible influence factors on prompt evaluation. Comparison prompts each varied in only one parameter. Hence, they were directly comparable for study participants with respect to their preferred manifestation of this parameter. Other than this modification, prompts remained in accordance with their comparison prompts in terms of general sentence structure, wording, and length. Modifications were conducted solely regarding one syntactical, grammatical, or lexical parameter at a time. In total, 1,044 prompts were designed and modified for the prompt types Functional Prompts, Informational Prompts, and Chit Chat Prompts. A prompt thereby qualified as a Functional Prompt when confirming the execution of a requested function, going beyond short responses such as ‘Okay’, ‘You got it’, etc. Querying user input also falls under this category. Informational Prompts are used to convey information about functions of the vehicle. The dialog is usually ended after one turn, but further communication is possible in form of Functional Prompts. Chit Chat Prompts are not linked to function execution but help users to get to know a VA better. The focus lies on the human not the assistant component of a VA with the goal oftentimes being to pass the time.

2.3 Conducting Exploratory Studies

Overall, six exploratory studies were conducted online via crowdsourcing in form of A/B testings with two studies for each Functional Prompts, Informational Prompts, and Chit Chat Prompts. Parameters were split randomly between the studies with each study containing between 84 and 90 prompt pairs. The first round of studies contained 11 parameters queried across Functional Prompts, Informational Prompts, and Chit Chat Prompts. The second study round incorporated the remaining 17 parameters, also querying them for Functional Prompts, Informational Prompts, and Chit Chat Prompts.

Participants obtained a general introduction to the study asking them to imagine driving a premium vehicle with a built-in VA they can ask questions, give commands, and make small talk with. Prompt pairs were then presented randomly to study participants to avoid sequence effects. Participants were asked for their subjective evaluation of a prompt altered in only one parameter at a time with the rest of a respective prompt remaining unchanged. For each prompt pair, participants received prior instructions putting the respective prompts in an in-car context. After reading the instructions as well as the prompt pair, participants had to select the prompt version they intuitively liked better. Figure 1 illustrates this exemplarily[5]:



Instruction | Imagine, you ask your voice assistant if you are still on the fastest route. Which of the following answers do you intuitively like better?

Option 1 – Active Voice | You are no longer on the fastest route. Should I select the recommended route? It will get you to your destination faster.

Option 2 – Passive Voice | You are no longer on the fastest route. Should the recommended route be selected? It will get you to your destination faster.

Figure 1: Study Design

Instructions varied between Functional Prompts, Informational Prompts, and Chit Chat Prompts but were kept consistent within these categories, only differing in the description of the use case. For Functional Prompts, participants were asked to imagine carrying out a function or receiving a function execution confirmation, e.g., starting navigation, or inquiring about a faster route. Instructions on Informational Prompts requested depicting situations where information on car functions like e.g., the Assisted Driving Mode is needed. Chit Chat Prompts' directives invited participants to envision situations entailing chatting with questions like e.g. 'Are you human?'.

Although conducted on paper, participants were encouraged to imagine their conversations happening via speech. Participants received the prompts in written form, not auditorily. This approach was selected to prevent possible interaction effects with varying quality of text-to-speech (TTS) outputs across prompts. Furthermore, a study by Stier et al. found that study participants were not able to distinguish between two auditorily presented prompt variants differing in one syntactical parameter. They conclude that users can certainly evaluate prompts intuitively, picking

[5] The English translation is provided to ensure understanding, but the study was conducted in German.

their favorite from two auditorily presented prompts, with the limitation of not knowing what they base their decision on [132]. The nature of the present study lies in an exploration of user preferences regarding different prompt variants though. Hence, it was important that participants can detect clear differences between prompts and rate them accordingly. Furthermore, variation of parameters in the present studies were – in parts – more minor than Stier et als.' parameter and hence more fleeting when presented auditorily. Therefore, the conclusion was made to conduct the study in written form.

2.4 Participants

A total number of 1,206 native German speaking participants took part in the study with 401 evaluating Functional Prompts, 401 evaluating Informational Prompts, and 404 evaluating Chit Chat Prompts. Age of participants ranged from 18 to 66 with a mean of 33.49 years and 54.1% identifying male, 44.8% identifying female, and 1.1% not disclosing their sex. Most participants indicated to drive 5,000 to 10,000 kms per year (34.4%), followed by 10,000-20,000 kms (28.2%), and 0-5,000 kms (20.8%). Most participants set their usage of in-car VAs to 'several times a week' (30.4%) with only 6.1% daily users and 22.5% never using them. 20.2% stated to employ their in-car VAs several times a month while 20.7% engaged with their VAs more rarely.

3 RESULTS

An exploratory approach to the study was necessary due to research in this field being, to date, scarce. As no a priori hypotheses were defined due to the studies' exploratory nature, results are attributed a preliminary character and are not evaluated statistically. Since the sample size was furthermore set to 200 participants per study to uncover potential effects and study participants' preferences, statistical calculations would not reveal reliable significances [23]. Results are therefore presented strictly descriptively.

Research Question I: *What is the entirety of syntactical, grammatical, and lexical parameters with a potential impact on prompt design?* The review of a German contemporary grammar revealed 28 parameters with at least two distinguishable manifestations and thus the potential to impact users' perception of a prompt. These parameters can be grouped into four parameters attributed to syntax, 13 parameters referring to grammar as well as 11 parameters regarding lexis. Table 1 presents an overview over all extracted parameters that were modified for exploratory studies, thereby answering research question I. Column 1 (Category) embeds a respective parameter in its linguistic group. Column 2 (Parameter) names the exact parameter while column 3 (Manifestations & Definition) reveals which manifestations of a parameter were modified, furthermore entailing definitions on parameters.

Table 1: Overview over Syntactical, Grammatical, and Lexical Parameters

CATEGORY	PARAMETER	MANIFESTATIONS & DEFINITION
SYNTAX	Sentence Structure	Hypotactical sentences: alternation of main clauses and subordinate clauses | Paratactical sentences: sequence of main clauses | Multiple compound sentences: main clauses with more than two dependent subordinate clauses
SYNTAX	Sentence Length	Short sentences: less than 11 words | Medium-length sentences: 11 to 15 words | Long sentences: more than 15 words
SYNTAX	Position of Subordinate Clauses	Prepositive subordinate clauses: subordinate clause preceding the main clause in a hypotactical sentence | Postpositive subordinate clauses: subordinate clause following the main clause in a hypotactical sentence | Interpositive subordinate clauses: subordinate clause located in the middle of the main clause in a hypotactical sentence
SYNTAX	Word Order	Word order: Subject – Predicate – Object | Word order: Object – Predicate – Subject | Word order: Predicate – Subject – Object Contrary to English, German allows for a more flexible word order without altering the meaning of a sentence.
GRAMMAR	Voice	Active voice | Passive voice
GRAMMAR	Main Verbs vs Auxiliaries	Use of main verbs | Use of auxiliary verbs
GRAMMAR	Verbal Style vs Nominal Style	Use of nominal style | Use of verbal style
GRAMMAR	Contractions	In German, articles and prepositions can be merged to a contracted form. Article-preposition-contractions are a characteristic trait of spoken language German [52]. Contracted article and preposition (e.g., German 'ans') | Separated article and preposition (e.g., German 'an das')
GRAMMAR	Present Participle: Attributive Adjectives	Use of a present participle construction in form of an attributive adjective | No participle construction
GRAMMAR	Present Participle: Subordinate Clauses	Sentence with a subordinate clause | Omission of a subordinate clause by usage of a present participle construction
GRAMMAR	Past Participle: Attributive Adjectives	Use of a past participle construction in form of an attributive adjective | No participle construction
GRAMMAR	Past Participle: Subordinate Clauses	Sentence with a subordinate clause | Omission of a subordinate clause by usage of a past participle construction
GRAMMAR	Grammatical Mood	Use of indicative | Use of subjunctive
GRAMMAR	Comparisons	Comparison with comparatives | Comparison with superlatives
GRAMMAR	Case	Use of genitive | Use of dative
GRAMMAR	Present vs Past Tense	Use of present tense (e.g., 'The temperature is already at 22 degrees.') | Use of past tense (e.g., 'The temperature was already at 22 degrees.')
GRAMMAR	Present vs Future Tense	Use of present tense (e.g., 'I'm starting the navigation right away.') | Use of future tense (e.g., 'I'll start the navigation right away.')
LEXIS	Interjections	Use of interjections (e.g., 'hm', 'whew') | No use of interjections

LEXIS	Conjunctions	Use of conjunctions to connect sentences | No use of conjunctions
LEXIS	Relative Pronouns	Informal relative pronouns: der/die/das | Formal relative pronouns: welcher/welche/welches
LEXIS	Ellipsis vs Repetition: Subject	Ellipsis of a mentioned subject | Repetition of a mentioned subject
LEXIS	Ellipsis vs Repetition: Verb	Ellipsis of a mentioned verb | Repetition of a mentioned verb
LEXIS	Ellipsis vs Repetition: Object	Ellipsis of a mentioned object | Repetition of a mentioned object
LEXIS	Referencing	Self-referencing of the VA: use of 1st person singular | Joint referencing of VA and driver: use of 1st person plural | Driver-only referencing: use of 3rd person plural
LEXIS	Adjectives	None: no adjectives | Low: adjectives make up 10% of a sentence | Medium: adjectives make up 25% of a sentence | High: adjectives make up 40% of a sentence
LEXIS	Adverbs	None: no adverbs | Low: adverbs make up 10% of a sentence | Medium: adverbs make up 25% of a sentence | High: adverbs make up 40% of a sentence
LEXIS	Polite Form	Level 0: no politeness markers | Level I: lexical politeness markers (‘please’, ‘upon request’) | Level II: grammatical politeness markers (through subjunctive constructions) | Level III: combination of lexical & grammatical politeness markers
LEXIS	Modal Particles	In German, modal particles serve as accentuation and determine the tone of voice. They do not contribute to a sentence semantically but act on an expressive level, able to transport speakers’ attitudes towards content [50]. None: no modal particles | Low: modal particles make up 10% of a sentence | Medium: modal particles make up 25% of a sentence | High: modal particles make up 40% of a sentence

Research Question II. *Which manifestation of syntactical, grammatical, and lexical parameters is preferred by participants for which prompt type?* All parameters in table 1 were subsequently fed into the exploratory studies described in chapter 2.3. The studies with a total of 1,206 participants evaluating 1,044 prompts revealed an impact of syntactical, grammatical, and lexical parameters on prompt perception. Results demonstrated participants’ susceptibility for different manifestations of parameters across varying prompt types. Outcomes displayed partly strong tendencies with up to 97% of participants agreeing on one manifestation of a parameter in direct comparisons as table 2 illustrates below. Column 1 embeds parameters in column 2 in their respective linguistic group. Columns 3 to 5 show test persons’ preferences for one or the other manifestation of a given parameter across the prompt types Functional Prompts (column 3), Informational Prompts (column 4), and Chit Chat Prompts (column 5). An example on how to read the table: the parameter ‘Sentence Structure’ is attributed to the linguistic group ‘Syntax’ and has three manifestations: hypotactical sentences, paratactical sentences, and multiple compound sentences. All three manifestations were compared amongst each other across the prompt types

Functional Prompts, Informational Prompts, and Chit Chat Prompts. The comparison of hypotactical sentences (abbreviated as 'Hypo') and paratactical sentences (abbreviated as 'Para') revealed a preference for hypotactical sentences with 68% of study participants favoring this manifestation for Functional Prompts. As described earlier, test persons' preferences for manifestations of a given parameter were tested by providing three pairs of prompt examples only differing in this respective parameter's manifestations. For each of these pairs, percentages for test persons' preferences were calculated. These values were then cumulated into one percentage value describing test persons' decisions regarding their preferred manifestation of one parameter overall[6]. Table 2 shows the respectively favored and disfavored options, marked in orange and green. Light green thereby means up to 74% accordance amongst participants for a given parameter's manifestation, medium green means between 75%-89% accordance, and dark green means between 90%-100% accordance. The color scheme for the respectively disfavored option in orange fits accordingly. Cells marked in light grey in column 2 show where preferences for parameters' manifestations were in accordance across Functional Prompts, Informational Prompts, and Chit Chat Prompts.

Table 2: Overview over Results

	Parameter	**Functional Prompts**						**Informational Prompts**						**Chit Chat Prompts**					
SYNTAX	Sentence Structure[7]	Hypo	Para	Hypo	MCS	Para	MCS	Hypo	Para	Hypo	MCS	Para	MCS	Hypo	Para	Hypo	MCS	Para	MCS
		68%	32%	81%	19%	68%	32%	68%	32%	73%	27%	50%	50%	73%	27%	66%	34%	57%	43%
	Sentence Length[8]	Short	Med	Short	Long	Med	Long	Short	Med	Short	Long	Med	Long	Short	Med	Short	Long	Med	Long
		39%	61%	55%	45%	69%	31%	19%	81%	23%	77%	37%	63%	26%	74%	47%	53%	81%	19%
	Position of Sub-Clauses[9]	Pre	Post	Pre	Inter	Post	Inter	Pre	Post	Pre	Inter	Post	Inter	Pre	Post	Pre	Inter	Post	Inter
		71%	29%	89%	11%	82%	18%	34%	66%	70%	30%	91%	9%	33.5	66.5	69%	31%	88%	12%
	Word Order[10]	SPO	OPS	SPO	PSO	OPS	PSO	SPO	OPS	SPO	PSO	OPS	PSO	SPO	OPS	SPO	PSO	OPS	PSO
		77.5	22.5	82.5	17.5	73%	27%	74%	26%	78%	22%	71%	29%	79%	21%	84%	16%	62%	38%
GRAMMAR	Voice	Active			Passive			Active			Passive			Active			Passive		
		90%			10%			58%			42%			91%			9%		
	Main Verbs vs Auxiliaries	Main Verbs			Auxiliary Verbs			Main Verbs			Auxiliary Verbs			Main Verbs			Auxiliary Verbs		
		82%			18%			82%			18%			61%			39%		
	Verbal Style vs Nominal	Verbal Style			Nominal Style			Verbal Style			Nominal Style			Verbal Style			Nominal Style		
		70%			30%			73%			27%			67%			33%		
	Contractions	Contracted			Separated			Contracted			Separated			Contracted			Separated		
		89%			11%			61%			40%			89%			11%		
		No Participle			Participle			No Participle			Participle			No Participle			Participle		

[6] All values are rounded.

[7] Hypo = hypotactical sentences, Para = paratactical sentences, MCS = multiple compound sentences

[8] Med = medium-length

[9] Pre = prepositive subordinate clause, Post = postpositive subordinate clause, inter = Interpositive subordinate clause

[10] SPO = subject, predicate, object, OPS = object, predicate, subject, PSO = predicate, subject, object

	Parameter	Functional Prompts						Informational Prompts						Chit Chat Prompts					
GRAMMAR	Present	91%			9%			87%			13%			65%			35%		
	Present Participle:	No Participle			Participle			No Participle			Participle			No Participle			Participle		
		53%			47%			43%			57%			76%			24%		
	Past Participle:	No Participle			Participle			No Participle			Participle			No Participle			Participle		
		54%			46%			77,5%			22.5%			66%			34%		
	Past Participle:	No Participle			Participle			No Participle			Participle			No Participle			Participle		
		51.5%			48.5%			42%			58%			51%			49%		
	Grammatical Mood	Indicative			Subjunctive			Indicative			Subjunctive			Indicative			Subjunctive		
		81%			19%			93%			7%			92.5%			7.5%		
	Comparisons	Comparative			Superlative			Comparative			Superlative			Comparative			Superlative		
		51%			49%			73%			27%			49%			51%		
	Case	Genitive			Dative			Genitive			Dative			Genitive			Dative		
		86.5%			13.5%			62%			38%			81.5%			18.5%		
	Present vs Past Tense	Present			Past			Present			Past			Present			Past		
		85%			15%			91%			9%			91%			9%		
	Present vs Future Tense	Present			Future			Present			Future			Present			Future		
		90%			10%			90%			10%			80%			20%		
	Interjections	No Interjections			Interjections			No Interjections			Interjections			No Interjections			Interjections		
		95%			5%			93%			7%			80%			20%		
L	Conjunctions	No Conjunctions			Conjunctions			No Conjunctions			Conjunctions			No Conjunctions			Conjunctions		
		71%			29%			57%			43%			73%			27%		
LEXIS	Relative Pronouns	Informal			Formal			Informal			Formal			Informal			Formal		
		82.5%			17.5%			78%			22%			84%			16%		
	Ellipsis vs Repetition:	Repetition			Ellipsis			Repetition			Ellipsis			Repetition			Ellipsis		
		80%			20%			29%			71%			68%			32%		
	Ellipsis vs Repetition:	Repetition			Ellipsis			Repetition			Ellipsis			Repetition			Ellipsis		
		31%			69%			21%			79%			47%			53%		
LEXIS	Ellipsis vs Repetition:	Repetition			Ellipsis			Repetition			Ellipsis			Repetition			Ellipsis		
		56%			44%			56%			44%			27%			73%		
	Referencing[11]	3rd pl	1st sg	3rd pl	1st pl	1st pl	1st sg	3rd pl	1st sg	3rd pl	1st pl	1st pl	1st sg	3rd pl	1st sg	3rd pl	1st pl	1st pl	1st sg
		52%	48%	89%	11%	13%	87%	90%	10%	91%	9%	51%	49%	69%	31%	63%	37%	69%	31%
	Adjectives[12]	None	Low	None	Med	None	High	None	Low	None	Med	None	High	None	Low	None	Med	None	High
		91%	9%	92%	8%	91%	9%	57%	43%	82%	18%	89%	11%	73%	27%	84%	16%	86%	14%
LEXIS		Low	Med	Low	High	Med	High	Low	Med	Low	High	Med	High	Low	Med	Low	High	Med	High
		90%	10%	93.5	6.5%	87%	13%	86%	14%	90.5	9.5%	87%	13%	85%	15%	86%	14%	85%	15%
	Adverbs	None	Low	None	Med	None	High	None	Low	None	Med	None	High	None	Low	None	Med	None	High
		91%	9%	95%	5%	96%	4%	77%	23%	83%	17%	85%	15%	71%	29%	81%	19%	83%	17%
		Low	Med	Low	High	Med	High	Low	Med	Low	High	Med	High	Low	Med	Low	High	Me	H
		94%	6%	94%	6%	91%	9%	81.5	18.5	84%	16%	83%	17%	83%	17%	81%	19%	8%	22%
	Polite Form	0	I	0	II	0	III	0	I	0	II	0	III	0	I	0	II	0	III
		53%	47%	87%	13%	85%	15%	85%	15%	94%	6%	92%	8%	65%	35%	83%	17%	86%	14%
LEXIS		I	II	I	III	II	III	I	II	I	III	II	III	I	II	I	III	II	III
		90%	10%	90%	10%	57%	43%	68%	32%	90%	10%	84%	16%	82%	18%	85%	15%	63%	36%
	Modal Particles	None	Low	None	Med	None	High	None	Low	None	Med	None	High	None	Low	None	Med	None	High
		92%	8%	97%	3%	93%	7%	90%	10%	95%	5%	92%	8%	32%	68%	85%	15%	83%	17%
		Low	Med	Low	High	Med	High	Low	Med	Low	High	Med	High	Low	Med	Low	High	Med	High
		94%	6%	96%	4%	97%	3%	94%	5%	95%	5%	96%	4%	52%	48%	65%	35%	76%	24%

[11] 1st sg = 1st person singular, 1st pl = 1st person plural, 3rd pl = 3rd person plural
[12] Med = Medium

Parameter	Functional Prompts	Informational Prompts	Chit Chat Prompts
Preferred manifestation in accordance with Functional Prompts, Informational Prompts, and Chit Chat Prompts	50%-74%	75%-89%	90%-100%
	25-50%	10-24%	9%-0%

Research Question III & Research Question IV. Which design guidelines and best practices can be distilled on syntactical, grammatical, and lexical levels? & Do results allow for identification of overall design patterns? Overall, results revealed not only compelling insights into best practices in prompt design, but also identified bigger picture user needs regarding system outputs. When interpreting study results, three main user needs emerged from observing overarching patterns in study participants' preferences regarding manifestations of parameters: a) a suitable level of (in)formality, b) a suitable level of complexity/simplicity, and c) a suitable level of (im)mediacy. A suitable level of formality is a cross of formal and less formal formulations. While a high level of formality oftentimes leads to less natural-language prompts, a high level of informality tilts a prompt too much towards colloquial language. Both 'extremes' were thereby rejected by study participants. Complexity/Simplicity refers to higher/lower cognitive demand for processing prompts and is oftentimes expressed on a syntactical level. Generally, simpler prompts were preferred by study participants. Finally, the term 'Immediacy' is used to describe the observed tendency of study participants to prefer straightforward formulations. Unnecessary information in form of linguistic bells and whistles was frequently penalized by test persons across prompt types and parameters. The fewer unnecessary information a prompt contains, the more immediate it is. These observations allow for establishing of the following three main guidelines for designing prompts:

1. Prompts should be written in natural and rather informal language without being too colloquial.
2. Prompts should be written in plain and simple language to avoid complexity.
3. Prompts should be written results- and information-oriented, leaving out unnecessary elements.

To fill these guidelines with concrete best practices, all parameters evaluated in the exploratory studies were assessed regarding their possible categorization under the defined user needs. Table 3 presents this categorization showing preferred manifestations of parameters in green and disfavored manifestations in orange in columns 2 and 3. Column 1 lists the parameters, with grey shaded cells marking parameters evaluated consistently across prompt types. Parameters were mapped to the user need of formality-informality if they have both a written-language and a spoken-language manifestation. Parameters fall under the user need of complexity-simplicity when they add to or reduce cognitive load in their different manifestations. The more unusual a formulation is, the more it thereby adds to cognitive load. Finally, parameters are considered part of the mediate-immediate user need if their manifestations are (non-) informational and add to, respectively hinder direct formulations. Parameters falling under (im)mediacy also impact a prompt's content: a prompt formulated in the present tense is more immediate than formulated in the past or future tense; a prompt formulated with an auxiliary verb is less immediate than one with a main verb.

Table 3: Best practices according to user needs

Parameter	**User need**	
	Formality	**Informality**
Case	Genitive	Dative
Grammatical Mood	Subjunctive	Indicative
Verbal Style vs Nominal Style	Nominal Style	Verbal Style
Voice	Passive Voice	Active Voice
Contractions	Separated Article and Preposition	Contracted Article and Preposition
Relative Pronouns	Formal: welcher/welche/welches	Informal: der/die/das
Interjections	No Interjections	Interjections
Comparisons	Comparative: FPs and IPs [a]	Superlative: CCPs[a][13] only
Polite Form	Lexical and Grammatical Politeness	No Politeness
Present Participle: Attr. Adj.	Use of Participles	No Use of Participles
Present Participle: Subord. Cl.	Use of Participles: IPs only	No Use of Participles: FPs and CCPs
Past Participle: Attr. Adj.	Use of Participles	No Use of Participles
Past Participle: Subord. Cl.	Use of Participles: IPs only	No Use of Participles: FPs and CCPs
Parameter	**Complexity**	**Simplicity**
Sentence Structure	Multiple Compound Sentences	Hypotactical Sentences
		Paratactical Sentences
Sentence Length	Long Sentences: IPs only	Short Sentences
		Medium-length Sentences: FPs and CCPs
Position of Subordinate Clauses	Interpositive Subordinate Clauses	Postpositive Subordinate Clauses: IPs and CCPs

[13] FPs = Functional Prompts, IPs = Informational Prompts, CCPs = Chit Chat Prompts

Parameter	User need	
		Prepositive Subordinate Clauses: FPs only
Word Order	Object-Predicate-Subject Predicate-Subject-Object	Subject-Predicate-Object
Parameter	**Mediacy**	**Immediacy**
Main Verbs vs Auxiliaries	Auxiliary Verbs	Main Verbs
Conjunctions	Conjunctions	No Conjunctions
Present vs Future Tense	Future Tense	Present Tense
Present vs Past Tense	Past Tense	Present Tense
Ellipsis vs Repetition: Subject	Repetition of Subjects	Ellipsis of Subjects: IPs only
Ellipsis vs Repetition: Verb	Repetition of Verbs	Ellipsis of Verbs
Ellipsis vs Repetition: Object	Repetition of Objects	Ellipsis of Objects: CCPs only
Adjectives	Adjectives	No Adjectives
Adverbs	Adverbs	No Adverbs
Modal Particles	Modal Particles: low amount for CCPs	No Modal Particles
Referencing	1st person plural	3rd person plural 1st person singular

This table shows the closing overview over all evaluated parameters and their preferred manifestations. It furthermore demonstrates how the presented guidelines can be implemented in terms of syntactical, grammatical, and lexical best practices.

4 DISCUSSION

Prior work has established that users are susceptible to differing prompt formulations on e.g. a syntactical level [132]. The research area of linguistic-centred prompt design is nonetheless far from being exhausted. Previous studies did not overarchingly cover a broad range of syntactical, grammatical, and lexical parameters important for prompt perception in a structured and methodical manner. Through review of a German contemporary grammar, 28 parameters with a potential impact on perception of prompts were compiled. As mentioned in chapter 2.1, some originally extracted parameters proved to be insufficient for the studies' purposes as they were not comparable across prompt types or – due to their inability to form two or more comparable manifestations – in general. While the examined parameters provide a comprehensive insight into best practices for prompt design in German, their expansion and refinement could be driven further in future studies to enhance their completeness. In addition, collecting qualitative data to learn about participants' reasons for preferring manifestations represents an interesting starting point for future studies as well.

The selected parameters were presented to participants in exploratory studies, asking them for their evaluation of prompts across different prompt types. In total, 1,206 study participants evaluated 1,044 prompts, differing in respectively one syntactical, grammatical, or lexical parameter. Results demonstrated that participants prefer certain syntactical, lexical, and grammatical manifestations over others. Preferences for parameters were in accordance across 67,86% of parameters, showing that design principles correlate over prompt types. Nonetheless, the number also demonstrates that – in parts – a different approach to prompt design needs to be explored for Functional Prompts, Informational Prompts, and Chit Chat Prompts. As the different prompt types cater to different user needs, this result was assumed. Users expect quick and direct function-related system outputs from Functional Prompts while they call for more diverting and social-driven conversations when chatting with a VA through Chit Chat Prompts. Informational Prompts address the users' need for information around car functionalities, thereby oftentimes replacing user manuals.

This paper's findings complement already existing VUI design guidelines regarding dialog guidance and dialog management and extend them with concrete instructions for prompt design on a linguistic level. In addition, results allowed for distilling three main user needs which cluster all evaluated parameters and enable the formulation of three linguistic-centered design guidelines. These guidelines are adaptable for VAs also beyond in-vehicle assistants. Guidelines as well as the parameters constituting them will be discussed in the following paragraphs.

4.1 Discussion of guidelines and parameters

Formality-Informality. When designing for VUIs, one design standard is formulating prompts in natural language, resembling spoken dialog [75, 108]. Results of the present studies mostly confirm this: of the 13 parameters attributed to the user need of formality-informality, 11 parameters were preferred in their informal and hence more natural-language manifestation. Active voice as well as verbal style were preferred over passive voice and nominal style. The latter two represent a more economical, compressed, and rational language used frequently in written contexts and in official German [131]. While study participants also preferred the active voice in Informational Prompts, results were less explicit compared to results for Functional Prompts and Chit Chat Prompts. When using an active grammatical voice, the agent is in the foreground whereas for passive grammatical voice, the action is [26]. A higher tolerance for passive voice in Informational Prompts – whose task it is to provide information about functions, therefore focusing on an action and less on an agent – is well understandable. On a further note regarding verbal style: research indicates that there seems to be gender differences in the use of nominal vs verbal style. Men tend to use nominal style, women by contrast verbal style [84]. This might be applicable for VAs which are most commonly equipped with female voices [23]. While natural-language and informal manifestations of parameters were mostly preferred by study participants, there are exceptions. Parameters with informal manifestations which are also connotated strongly colloquially were rated worse than their formal counterparts. An explanation can be found in that '[t]he very nature of being a computer may limit its ability to appropriately and capably employ certain linguistic concepts that are inherently social' as proposed by Clark [32]. In this study, participants strongly penalized the use of interjections in prompts, thereby supporting Clark's theory of social boundaries in HCI as interjections too are a typical means of social and colloquial HHI. Additionally, interaction effects with the prompt type were observed for some parameters. For Informational Prompts, study participants preferred the respectively less natural-language manifestation of the parameters comparisons and participles (for 'Participles: Subordinate Clauses'). This is particularly interesting as Informational Prompts oftentimes emerge from written instruction manuals, potentially leading to higher tolerance for written-language linguistic characteristics on user side. Lastly, respectively less polite and hence more informal prompt versions were preferred by study participants across all three prompt types. In line with this result, the informal indicative was also favored over the more formal subjunctive. Politeness is an important factor in prompt design and can be an especially essential parameter for error handling and conversational repair structures. Intuitively, when designing for a service-driven VA, a high level of politeness seems like the right design choice. Especially when employing politeness in form of subjunctives, politeness gives interactions a non-committal appearance by toning down the demanding nature of a statement [149]. Again, Clark can be cited, suggesting that a high level of politeness as a naturally social linguistic concept in HHI can lead into the uncanny valley when applied in HCI [32].

Complexity-Simplicity. A prompt's complexity is oftentimes manifested on a syntactical level. The more complex a syntactical composition, the more complex production and perception are too [26, 44]. In line with findings from Stier et al. and Demberg et al., complexity in form of intricate syntactical structures was penalized by participants of our studies [38, 132]. Across prompt types, participants preferred the most commonly used and hence least complex word order of German [81, 131], namely subject-predicate-object. Study participants also penalized too complex sentence structures as in multiple compound sentences, thereby being consistent with Chafe and Stier [26, 132]. They did prefer slightly more complex hypotactical sentences over paratactical sentences though. A possible explanation is that the paratactical concatenation of more sentences leads to telegram style, neither sounding natural nor like a modern-age VA's output: prompts with four distinct paratactical sentences went along with a worse evaluation than prompts with three distinct sentences. Results for sentence length varied between prompt types. For Informational Prompts, long prompts (more than 15 words) were preferred over short prompts (less than 11 words), and medium-length prompts (11-15 words). As the focus of Informational Prompts is conveying information, a higher tolerance for an increased sentence length does not come unexpected. For Chit Chat Prompts and Functional Prompts, medium-length sentences were the preferred sentence length. Chit Chat Prompts and Functional Prompts cater to a different user need than Informational Prompts. Chit chat is usually invoked by users to fulfil the user need of conversing on a social level. Functional Prompts on the other hand correlate with a different sense of immediacy as they are a means to carry out functions and get support with a specific task. The basic need is rather task- than communication-oriented and possibly time-critical, potentially resulting in lower tolerance for longer prompts. Furthermore, results suggest that a prompt becoming longer through addition of more words with informational character is rather accepted than comparison prompts increasing in length through adding non-informational adjectives, adverbs, or modal particles. Length of prompts is certainly a complex parameter with dependencies on attention a given user can spare at a given moment and individual cognitive processing capabilities [1]. Design guidelines for sentence length are therefore delicate to generalize in theory as road conditions, cognitive load, and respective cognitive processing capabilities can differ broadly in practice.

Mediacy-Immediacy. In our studies, participants preferred straightforwardly formulated prompts. Of 11 parameters attributed to the user need of mediacy-immediacy, only three parameters were preferred in their more mediate manifestation. Results for these parameters furthermore varied across prompt types. Immediacy can be expressed through avoiding unnecessary linguistic elements in form of repetitions and non-informational adjectives, adverbs, or modal particles. A low number of modal particles (modal particles make up 10% of a sentence) was preferred over no modal particles for Chit Chat Prompts. For all other comparisons and across prompt types, the prompt example with the respectively fewer number of adjectives, adverbs, and modal particles was preferred by study participants. Except for the small tolerance for Chit Chat Prompts where a VA's social component – which is neither information-driven nor function-dependent – is in the focus, this shows study participants' low tolerance towards linguistic elements not contributing to a prompt on an informational level. Directness and clarity also constitute an immediate prompt. Here, immediacy is expressed through choice of tense or form of address (e.g., referencing or use of main verbs vs auxiliaries). Prompts written in future or past tense were rated worse than their comparison prompts in the present tense. Also, prompts with auxiliary verbs were less popular than comparison prompts with a main verb. These outcomes show study participants' wish for a direct and immediate form of communication. This wish is also mirrored in results for the parameter 'Referencing'. Across all prompt types, study participants' preferred form of address was 3rd person plural ('you'/German 'Sie'). Self-referencing of the system ('I'/German 'Ich) and joint referencing of the personal assistant and the user ('we'/German 'wir') were the respectively disfavored options. A possible explanation is that a direct form of address through 3rd person plural may counteract the issue of a diminished sense of agency when interacting with a speech interface as discussed in Limerick et al. [83]. According to Watzlawick, communication can be either complementary or symmetrical. Symmetrical communication includes communication with partners and friends on the same hierarchical level. Complementary communication means the opposite and is comprised of e.g. communication with a superior or employee in a formal work environment [142]. Preferred referencing may be dependent on how communication with a respective VA is perceived. A more complementary communication with e.g., a service-driven assistant could penalize the usage of 1st person plural and even 3rd person plural. Results of the present study are consistent with this line of thought. Usage of 'I', focusing on what the system can do for the driver, was more popular than joint referencing using 'we' for Functional Prompts. For Chit Chat Prompts and Informational Prompts, 'we' was preferred over 'I' though, indicating a more symmetrical communication.

4.2 Limitations

Prompts were queried in an online survey with stimuli being presented on paper, not auditorily. This approach is justified in light of study findings by Stier et al. [132] and appropriate in regard to the goals of the study. Nonetheless, auditory presentation of prompts and interaction effects with TTS outputs may have an influence on evaluation of parameters. While study participants were asked to imagine an in-car setting and evaluated actual in-vehicle prompts, results were not obtained under real driving conditions and in direct interaction with a VA. However, both factors represent considerable contexts. Hence, following up on the present results in driving simulator studies is necessary to determine the impact of a) the interaction of prompt preferences with the actual in-car VA performance (including e.g., TTS, ambient noises, multi-step dialogues, and multimodality) and b) interaction of prompt preferences with the primary driving task (considering i.e., cognitive load and multitasking scenarios). Further interaction effects could be expected from within parameters themselves (with e.g. sentence length), different contexts and use cases (e.g. in-car vs. smart home assistant), and from user side as indicated by current research [134]. These constraints should therefore be followed up on and addressed in future studies. At this point, it is essential to emphasize that results presumably are language-dependent and that findings of this study are primarily applicable for German. Yet, participants' susceptibility towards variation of parameters suggests effects in other languages as well. Furthermore, the extracted parameters can pose guidance in regard to which parameters to select for comparable studies in other languages.

5 CONCLUSION

This paper's contribution is twofold: for one, exploratory studies with a total of 1,206 participants evaluating 1,044 prompts revealed syntactical, grammatical, and lexical best practices for formulating prompts across different prompt types. Secondly, the interpretation of these study outcomes allowed for identification of three main user needs, which are incorporated into the following guidelines:

1. Prompts should be written in natural and rather informal language without being too colloquial.
2. Prompts should be written in plain and simple language to avoid complexity.
3. Prompts should be written results- and information-oriented and leave out unnecessary elements.

While these guidelines may seem intuitive, their composition and the syntactical, grammatical, and lexical best practices which constitute them, were to date not supported methodically. To our knowledge, these results thus pose a novelty as they represent the first comprehensive and methodical linguistic-based design guidelines for VAs on syntactical, grammatical, and lexical levels. These guidelines are applicable for a broad range of VA settings. By applying suitable manifestations of parameters, prompts can be tailored to fit to individual user needs in terms of (in)formality, complexity/simplicity, and (im)mediacy. Moreover, the results represent a sound foundation for future research in natural language generation, best practices in prompt design for further language families, and natural speech vs TTS. One must not only focus on dialog management and dialog guidance when designing a VUI, but also consider the linguistic level of prompts as it has an influence on user experience. More attention must be paid to these linguistic aspects to design VUIs in a user-centric way. This paper closes the gap of insufficient guidelines and lack of methodical research in the field of linguistic-centric prompt design and provides a valuable handbook on how to design prompts across prompt types.

III. SECURE, COMFORTABLE OR FUNCTIONAL: EXPLORING DOMAIN-SENSITIVE PROMPT DESIGN FOR IN-CAR VOICE ASSISTANTS[14]

User Experience in Human-Computer Interaction is composed of a multitude of building blocks, one of which is how Voice Assistants (VAs) talk to their users. Linguistic considerations around syntax, grammar, and lexis have proven to influence users' perception of VAs. Users have nuanced preferences regarding how they want their VAs to talk to them. Previous studies have found these preferences to differ between domains, but an exhaustive and methodical overview is still outstanding. By means of an A/B study spanning over domains as well as dialog types, this paper methodically closes this gap and explores the degree of domain-sensitivity across different types of dialogs in German. The results paint a mixed picture regarding the importance of domain-sensitivity. While some degree of domain-sensitivity was found for in-car prompts, it generally seems to play a rather minor role in users' experience of VAs in the vehicle.

1 INTRODUCTION

How Voice Assistants (VAs) communicate with their users comprises an integral part of a strong user experience. Besides the possibility to positively influence user experience by means of a VAs text-to-speech voice [61, 100], concrete prompt formulations additionally affect the perception of VAs [94, 134]. Studies have shown specific and fine-grained user preferences for certain formulations on syntactical, grammatical, and lexical levels. Research furthermore found that these preferences differ between domains (meaning conversational contexts like e.g., security-relevant vs comfort-oriented). In more detail, Stier et al. identified different preferences for sentence structures when comparing dialogs around comfort functions and driving assistance functions [132]. Additionally, Meck et al. found the current dialog type and the concomitant conversational need between a given user and their VA to influence formulation preferences [94]. In their research, they contrast strictly functional conversations against conversations around information retrieval and chit chat scenarios. Like Stier et al. [132], they find syntactical preferences to depend on the type of dialog. Both Stier et al.'s [132] and Meck et al.'s [94] studies point to context-sensitive prompting being an important step towards increasingly conversational HCI, which users envision

[14] As the leading author, I developed the research idea as well as the experiment design and conducted and analyzed all described studies.

in conversations with “Perfect Voice Assistants” [141, p. 1]. This paper wants to combine Stier et al.’s and Meck et al.’s approaches and further engage in research around formulation preferences for different domains across dialog types, more specifically for in-car scenarios. Careful considerations around prompt formulations – while overall central in HCI – are especially crucial in the car as this environment needs to account for a security-relevant primary task: driving [135].

Based on Stier et al.’s [132] research around domains, three different domains will be explored: a) a security-relevant domain, b) a comfort-oriented domain, and c) a general functional domain. Dialogs in the functional domain are utilized to execute functions in a service-oriented and task-related manner. Dialogs in the security-relevant domain deliver specific and clear-cut information on potentially critical situations, while dialogs in the comfort-related domain are focused on users’ well-being and convenience. Originating from Meck et al.’s [94] considerations, formulation preferences for the aforementioned domains are examined across functional dialog, informational dialog, and proactive dialog. To analyze potential differences in formulation preferences, exploratory A/B crowdsourcing studies are conducted to answer the following research question: *Which effect do domains have on the preference for linguistic parameters across dialog types?* Results obtained from the exploration can be harnessed by HCI practitioners designing context-sensitive, conversational VA prompts.

2 RELATED WORK

Although general recommendations and design guidelines for VAs date back several decades [39, 101, 102, 108, 129], research has only recently delved into the research area of linguistic prompt design [94, 132, 134]. While general recommendations on how VA prompts are to be formulated exist, these recommendation seldomly go into detail on how they can be adhered to on a linguistic level. For instance, Alvarez et al. suggest designing prompts in “short and clear segments” [1, p. 157] but do not go into detail on how brevity and clarity can be achieved. Further general recommendations are delivered by Semmens et al. and Schmidt et al. who state that prompts need to be natural, precise, and thoroughly formulated [124, 127]. While these recommendations seem fitting and intuitive, they leave open how exactly they take shape.

Evidence of the importance of a well thought out prompt design is produced by Stier et al. though [132]. The researchers examined syntactical structures and their influence on driving performance and found a significant interaction between both factors. In a comparison of more complex and nested vs simpler and more linear sentence structures, the former led to an increase in drivers' cognitive load. The formulation of in-vehicle prompts is hence potentially safety-critical and can be controlled by adhering to appropriate sentence structures. Research by Haas et al. add considerations around prompt length: while being rated as useful and likeable as long prompts, short prompts were rated as more efficient than long prompts [51]. Stier et al. [132] yield further insights into best practices for designing prompts on syntactical levels by adding considerations around different VA domains. In their research, they found formulation preferences to significantly depend on the current domain. While paratactical sentences were preferred for conversations around driving assistance functions, hypotactical sentences achieved higher ratings in comfort-oriented dialogs. Next to Stier et al.'s [132] research, concrete linguistic-driven design guidance on how (in-car) prompts can be structured and formulated can be drawn from Meck et al. [94]. Other than Stier et al. [132], the researchers did not focus on VA domains but on different types of conversations. They compared functional dialogs (concerning function execution), information-centred dialogs (concerning the disclosure and distribution of information), and chit chat dialogs (small talk) across 28 syntactical, grammatical, and lexical categories. Study participants' formulation preferences for prompts partly differed between dialog types. Exemplarily: an active voice was preferred by study participants for functional prompts and chit chat prompts, while no significant tendency was displayed in informational prompts. Passive sentence structures emphasize an action rather than an agent. In both service-oriented functional and agent-centred chit chat dialogs, an active agent is more important than in information-centred dialogs, where emphasis rather lies on an action than on the agent itself. While research has shown the importance of prompt design on a linguistic level, it needs mentioning that factors such as a fitting synthetic voice [55, 73, 91], agent embodiment [77], and lexical user-computer alignment [85] are also crucial factors for a well-rounded user experience.

The previous paragraphs establish the importance of domain-sensitive prompt design and show the influence of dialog types on formulation preferences. This paper wants to combine both research directions and methodically examine formulation preferences between domains across different types of conversations.

3 METHOD

Two crowdsourcing studies were developed and conducted online to gain insights into formulation preferences across domains and dialog types. Following Meck et al., functional as well as informational dialogs were chosen as dialog types [94]. As proactivity is a most interesting use case in in-vehicle environments [124], proactive dialogs were selected as a third type of dialog. The selection of study domains was based on Stier et al. [132], who examined and found different formulation preferences for security-relevant and comfort-oriented domains. A third domain, namely a general functional domain was chosen for this study to extend Stier et al.'s considerations.

3.1 Study I: Validation of Domains

A preliminary online crowdsourcing study was conducted to determine representative use cases for each of the abovementioned domains. In a within-subjects single-choice task, 200 study participants were asked to map nine car functions onto one of the three domains a) security-relevant, b) comfort-oriented, and c) functional. The car functions are listed in Table 1 below. Each function was accompanied by a short explanatory text, disclosing the functions' application in the car.

Table 1: Overview over Study Use Cases

Remaining Range	Speed Control	Digital Key
Relaxing Mode	Climate Settings	Massage Function
Parking	Navigation	Calling

In a next step, prompts were developed for each domains' most representative use case and systematically varied regarding syntactical, grammatical, and lexical parameters. These prompts were then used as study prompts in the subsequent second crowdsourcing study.

3.2 Study Parameters and Comparison Prompts

As previous studies found formulation preferences on syntactical, grammatical, and lexical levels [94, 132], all three linguistic dimensions are considered in the present paper. The selection of parameters was based on Meck et al., who explored a total of 28 parameters [94]. As the present study is intended as an exploratory study probing the potential domain-specific occurrence of best practices for formulating in-car prompts, only a subset of Meck et al.'s parameters was selected as study parameters. Table 2 gives an overview over study parameters, compared manifestations, and examples.

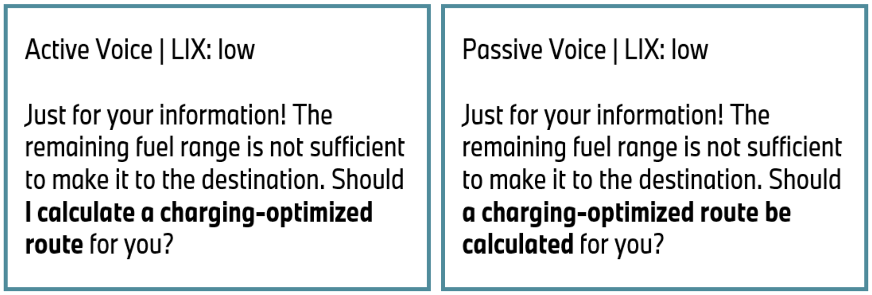
Table 2: Overview of Study Parameters & Example Prompts

<table>
<tr><th colspan="6">Sentence Structure[15]</th></tr>
<tr><td colspan="2"><i>Parataxes</i></td><td colspan="2"><i>Hypotaxes</i></td><td colspan="2"><i>Multi-Clauses</i></td></tr>
<tr><td colspan="2">Assisted Driving can help you with speed and lane keeping. It can also give you a better overview of your route. You do have the function in your car. Do you want to activate it?</td><td colspan="2">Assisted Driving cannot only help you with speed and lane keeping but it can also give you a better overview of your route. You do have the function in your car. Do you want to activate it?</td><td colspan="2">Assisted Driving cannot only help you with speed and lane keeping but it can also give you a better overview of your route and you do have the function in your car. Do you want to activate it?</td></tr>
<tr><th colspan="6">Sentence Length</th></tr>
<tr><td colspan="2"><i>Short</i></td><td colspan="2"><i>Medium</i></td><td colspan="2"><i>Long</i></td></tr>
<tr><td colspan="2">Assisted Driving helps with speed and lane keeping and can give you route overviews. Should I activate Assisted Driving?</td><td colspan="2">Assisted Driving can help you with speed and lane keeping and can also give you route overviews. Should I activate Assisted Driving for you?</td><td colspan="2">Assisted Driving can not only help you with speed and lane keeping but can give you a better overview of your route. Should I activate Assisted Driving for you?</td></tr>
<tr><th colspan="6">Form of Address</th></tr>
<tr><td colspan="2"><i>“I”</i></td><td colspan="2"><i>“you”</i></td><td colspan="2"><i>“we”</i></td></tr>
<tr><td colspan="2">Should I activate Assisted Driving?</td><td colspan="2">Do you want to activate Assisted Driving?</td><td colspan="2">Should we activate Assisted Driving?</td></tr>
<tr><th colspan="6">Position of Sub-Clauses</th></tr>
<tr><td colspan="3"><i>Prepositive</i></td><td colspan="3"><i>Postpositive</i></td></tr>
<tr><td colspan="3">If you activate Assisted Driving, I can automatically look for the fastest route.</td><td colspan="3">I can automatically look for the fastest route if you activate Assisted Driving.</td></tr>
<tr><th colspan="6">Politeness</th></tr>
<tr><td colspan="3"><i>With Lexical Politeness</i></td><td colspan="3"><i>Without Lexical Politeness</i></td></tr>
<tr><td colspan="3">I can gladly activate Assisted Driving for you.</td><td colspan="3">I can activate Assisted Driving for you.</td></tr>
<tr><th colspan="6">Voice</th></tr>
<tr><td colspan="3"><i>Active</i></td><td colspan="3"><i>Passive</i></td></tr>
<tr><td colspan="3">Should I activate Assisted Driving for you?</td><td colspan="3">Should Assisted Driving be activated for you?</td></tr>
</table>

Comparison prompts were designed for each parameter and its respective manifestations. The comparability and intelligibility of study prompts was ensured by calculating the so-called LIX index [82]. The LIX examines complexity and comprehensibility of text, by considering its number of words, number of clauses, the average clause length, and the number of long words (words with more than 6 characters). Only prompts with an equal LIX index were used as comparison prompts in the second study. Study prompts differed in one syntactical, grammatical, or lexical parameter at a time while the rest of a prompt was kept consistent. An exemplary comparison prompt can be found below in Figure 1. Two manifestations are compared for the parameter voice, namely active

[15] The study was conducted in German, but prompts are translated to English to broaden intelligibility.

voice, and passive voice. Comparison prompts for both manifestations are designed and examined regarding their LIX indices. As LIX is found to be low for both prompts, they are approved as study prompts.



Active Voice | LIX: low

Just for your information! The remaining fuel range is not sufficient to make it to the destination. Should **I calculate a charging-optimized route** for you?

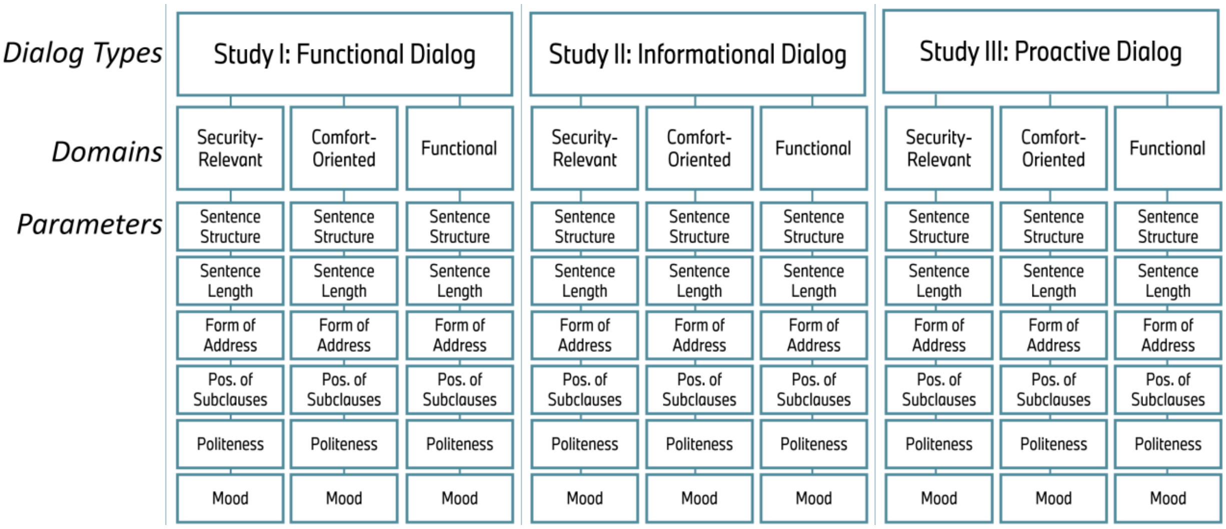
Passive Voice | LIX: low

Just for your information! The remaining fuel range is not sufficient to make it to the destination. Should **a charging-optimized route be calculated** for you?

Figure 1: Exemplary Study Prompts

3.3 Study II: Comparison of Formulation Preferences

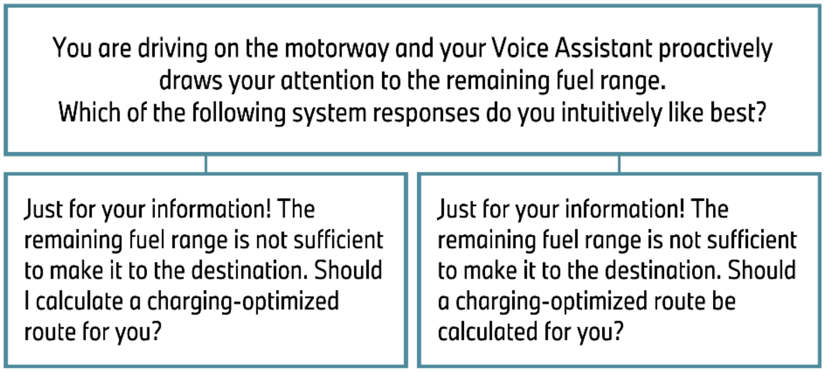
To explore formulation preferences across domains and dialog types, three online crowdsourcing studies were set up in a mixed factorial design and completed by 200 study participants each. The variable *dialog type* was altered between subjects and studies, while the variables *domains* and *parameters* were treated as within-factors. With this distribution, each study participant processed 54 comparison prompts. Figure 2 shows the studies' distribution and structure. Study prompts were presented to study participants in an A/B format which has proven to successfully detect differences between formulation preferences in previous studies [94]. Study participants received prompts in text form, not auditorily. The reasoning behind this decision is as follows: the fleeting nature of speech impedes the auditory comparison of two prompts which partly only differ in nuances [145]. Secondly, and more importantly, a study by Meck et al. found no differences in the evaluation of prompts between an online text-based crowdsourcing study and a driving simulator study, with auditorily presented study prompts [93]. Hence, it is argued that conclusions for actual in-car preferences can be drawn from text-based A/B crowdsourcing studies.



Dialog Types	Study I: Functional Dialog			Study II: Informational Dialog			Study III: Proactive Dialog		
Domains	Security-Relevant	Comfort-Oriented	Functional	Security-Relevant	Comfort-Oriented	Functional	Security-Relevant	Comfort-Oriented	Functional
Parameters	Sentence Structure	Sentence Structure	Sentence Structure	Sentence Structure	Sentence Structure	Sentence Structure	Sentence Structure	Sentence Structure	Sentence Structure
	Sentence Length	Sentence Length	Sentence Length	Sentence Length	Sentence Length	Sentence Length	Sentence Length	Sentence Length	Sentence Length
	Form of Address	Form of Address	Form of Address	Form of Address	Form of Address	Form of Address	Form of Address	Form of Address	Form of Address
	Pos. of Subclauses	Pos. of Subclauses	Pos. of Subclauses	Pos. of Subclauses	Pos. of Subclauses	Pos. of Subclauses	Pos. of Subclauses	Pos. of Subclauses	Pos. of Subclauses
	Politeness	Politeness	Politeness	Politeness	Politeness	Politeness	Politeness	Politeness	Politeness
	Mood	Mood	Mood	Mood	Mood	Mood	Mood	Mood	Mood

Figure 2: Distribution and Structure of Crowdsourcing Studies

Figure 3 illustrates the A/B structure of the study further. Study participants received a short text, introducing them to the use case and the potential in-car scenario. This text was followed by the comparison prompts, which only differed in one parameter at a time. Prompts were presented in randomized order to avoid sequence effects. Participants were asked to select the prompt they intuitively liked best.



You are driving on the motorway and your Voice Assistant proactively draws your attention to the remaining fuel range.
Which of the following system responses do you intuitively like best?

Just for your information! The remaining fuel range is not sufficient to make it to the destination. Should I calculate a charging-optimized route for you?

Just for your information! The remaining fuel range is not sufficient to make it to the destination. Should a charging-optimized route be calculated for you?

Figure 3: A/B Study Structure

4 RESULTS

4.1 Crowdsourcing Study

200 study participants were asked to map nine in-car use cases onto the three domains 1) security-relevant, 2) comfort-oriented, and 3) functional. 52% of study participants identified as male, 47% as female, and 1% as diverse. Average age of participants was 32.24 years (sd 9.73). Crowdworkers were obligated to possess valid driving licenses and to drive at least on a weekly basis. Furthermore, experience with in-car VAs was required. The majority of crowdworkers indicated to use their in-car VA every ride (42.5%), followed by every second ride (25.5%) or less regularly (26.4%). Solely 5.5% pointed out to have used their in-car VA only once.

A driving assistance function was rated highest in terms of security relevance (65.5%), while a relaxing mode was deemed most representative for the comfort-related domain (82.5%). Information on the remaining fuel range was found to best represent the functional domain (56.5%).

4.2 Formulation Preferences

A total of 600 study participants, distributed over three studies, rated 54 functional, informational, and proactive VA prompts to determine domain-sensitive best practices for in-car prompt formulations. The average age of study participants was 32.5 years (sd 10.04). 46.75% of study participants identified as female, 52.67% as male, and 0.58% as diverse. Participants were required to possess a valid driving license and to drive on at least a weekly basis. Moreover, experience with in-car VAs was obligatory. 41% of crowdworkers stated to user their in-car VA during every ride, while 24.5% indicated to use it every second ride. 29.5% of participants use their VAs less regularly, while 5% declared to only have used it once.

Due to the ordinal nature of the data, Kruskal Wallis tests were calculated in R [116]. No differences in formulation preferences were found for functional and informational prompts. Within proactive prompts, preferences for the position of sub-clauses differed significantly between domains: Chi-squared=6.4, df=2, p=0.04. Subsequent post-hoc Dunn Bonferroni tests showed a significant difference between the comfort-oriented and the functional domain (p=0.03), although the effect size was found to be small, Cohen’s r= 0.007.

5 DISCUSSION AND CONCLUSION

Within functional and informational prompts, no domain-specific formulation preferences were detected. Proactive prompts on the other hand called for domain-sensitive considerations around the position of sub-clauses. As such, the research question, namely "*Which effect do domains have on the preference for linguistic parameters across dialog types?"* – is arguably best answered with "*a small one*". While postpositive sub-clauses were preferred in the comfort-oriented domain, study participants preferred prepositive sub-clauses in functional dialog. Sub-clauses describe a main clause further and their position is important in terms of information packaging and information processing. In conversations, interlocutors package information in a manner that is most appropriate for the context and most easily processible by their conversational partner [26]. Sub-clauses are introduced with conditional, temporal, or causal conjunctions, which contain crucial information regarding what, when, why, or how something is happening in a main clause. By means of these conjunctions, prepositive sub-clauses directly indicate the reason for a proactive interruption [128], which is arguably more important in functional than in comfort-related dialog.

While a multitude of linguistic parameters were considered in the present study, it cannot raise a claim to completeness. Furthermore, the study was conducted in German and findings may be language and culture dependent. As the study only comprised in-car use cases, findings may not be applicable to other environments like e.g., the smart home. Future work could tend to find formulation preferences for VAs in these environments and furthermore concentrate on user characteristics, such as age, gender, or previous experience with VAs.

The results obtained in this exploratory study paint a mixed picture regarding the importance of domain-sensitive prompt design. Domain-specific formulation preferences were found for proactive dialogs, underlining the importance of linguistic considerations when designing prompts in HCI. On the other hand, and especially given how many individual parameters were assessed in the study, only one concrete domain-specific best practice emerged. In sum, while some degree of domain-sensitivity was found for in-car prompts, it generally seems to play a more minor role in users' experience of VAs in the vehicle. Rather, prompt preferences can be viewed as being more than the sum of their parts, as they depend on a multitude of factors, including the type of conversation and the agent presenting them.

IV. HOW MAY I INTERRUPT? LINGUISTIC-DRIVEN DESIGN GUIDELINES FOR PROACTIVE IN-CAR VOICE ASSISTANTS[16]

Voice Assistants are predicted to develop from merely reactive to increasingly proactive agents in the future. While proactivity allows a leap towards more intelligent conversations with Voice Assistants, designing proactive agents is not straightforward, as benefit and acceptance of proactive behavior is dependent on a plethora of factors. For instance, proactive agents run the risk of disrupting users who are already engaged in ongoing primary tasks. While a large body of research is therefore concerned with when to proactively interrupt users, how to interrupt them has received less attention. To close this gap, a driving simulator study was conducted to find linguistic best practices for designing proactive prompts in German. Low linguistic complexity as well as suggestive rather than imposing language were found to influence study participants' preferences for proactive prompts. These findings underline that the existing framework for designing proactive interactions needs to be enhanced by nuanced linguistic considerations.

1 INTRODUCTION

The past years have seen a rising interest around proactive Voice Assistants (VAs) from both industry and academia, as assistants are believed to develop from merely reactive to increasingly proactive agents [95, 104, 114, 118, 122, 126]. Proactivity entails system-initiated and largely autonomous VA behavior [104], which overcomes the prevalent "pull paradigm" [127, p. 2], where interactions emerge solely from the users' side. Contrary to reactive agents, proactive agents propose contextually relevant suggestions to users which can be utilized to anticipate problematic situations, prevent negative experiences, and furthermore enhance user experience altogether. Proactive behavior thus expands interaction possibilities in Human-Computer Interaction (HCI), which is in line with envisioned "perfect voice assistants" [141, p. 1] as proposed by Völkel et al., which are smart, personalized, and proactive. Proactivity hence allows a leap towards more natural as well as conversational and intelligent conversations with VAs, with use cases ranging from receiving recommendations or reminders, to helping with home safety and security, and even mental well-being [29, 80]. In the automobile, proactivity can be exceedingly helpful and provide

[16] As the leading author, I developed the research idea as well as the experiment design and conducted and analyzed all described studies. Dr. Christoph Draxler and Dr. Thurid Vogt supported the development of the research idea and the experiment design and provided feedback on the overall work.

drivers with information around i.e., the remaining fuel range or upcoming service needs. Driving scenarios are found to provide a rich environment for proactive interactions [127] and proactive behavior is found to be rated as positively as reactive behavior and to receive high acceptance rates, as well as high additional value scores [122, 124].

Although rated generally positively by users, proactivity is not without controversy and designing proactive use cases is not always straightforward [122, 124, 148]. Zargham et al. even speak of a “proactivity dilemma” [148, p. 1] as the benefit and acceptance of proactive behavior is dependent on a multitude of factors. Proactive agents run the risk of disrupting users who are already engaged in ongoing (social) interactions or conduct (potentially even security-relevant) primary tasks. Not solely but partly because of the afore-mentioned reason, proactive suggestions need to be contextually relevant. In case of in-car interactions, driving-related use cases were found to achieve higher user ratings than comfort-related use cases [123]. Zargham et al. [148] mirror this finding in that they report that situational more relevant use cases received higher acceptance ratings in a study they conducted. Moreover, the style a proactive agent addresses users in plays an essential role in successful proactive interactions. Reicherts et al. [114] find users to differentiate between suggested and imposed proactive behavior and to prefer the former. Content, timing, and style can hence be identified as defining factors regarding the suitability of proactive suggestions. A negative example which is frequently quoted in this context is Microsoft’s discontinued proactive office assistant Clippit. Clippit interrupted users intrusively (style) with nonessential information (content) and oftentimes during task execution (timing). Although not a voice-based agent, lessons for a strong user experience can be drawn from Clippit.

Regarding appropriateness of proactive interactions, the vehicle represents a particularly delicate environment. VA users are consistently occupied with a highly engaging and security-relevant primary task: driving. Schmidt et al. [122–124] conducted extensive research around proactivity in the car and found that proactivity may not be obtrusive and overload drivers. Drivers are known to react to secondary tasks with compensatory driving behavior and to reduce speed and micromanage the position of their steering wheel [62]. Alternatively, they lower their secondary task engagement to focus on the driving task [63], meaning primary and secondary tasks are highly intertwined in automotive settings and proactive interactions can potentially compromise driving safety. Research around proactive in-car interactions therefore strongly focusses on timing of interactions, so-called “opportune moments”. Moments are thereby considered opportune if drivers are currently not too preoccupied with their primary driving task [120, 127].

Successful proactive interactions are highly dependent on the factors primary task engagement, opportune moments, and interruptibility. These factors constitute *when* to interrupt users proactively. An equally fruitful body of research can be drawn upon regarding proactive contents, so *what* to proactively suggest to users, for both inside and outside in-vehicle scenarios [114, 124, 148]. *How* to interrupt users on the other hand has not been studied in detail yet. Research has been conducted as to whether users prefer to be addressed proactively by means of earcons and sound or by voice [56, 114]. However, concrete formulations of proactive prompts have not yet been examined. Research by Stier et al. [132] and Meck et al. [94] shows that users have particular preferences regarding the formulation of prompts though – both on syntactical, grammatical, as well as lexical levels. Their research finds that prompt formulations can be a decisive factor for a well-rounded user experience. Nonetheless, this research has overlooked proactivity so far. Reicherts et al.’s [114] finding that proactive suggestions rather than impositions are preferred by users provides a starting point for linguistic considerations around suggestive/imposing language. Schmidt et al.’s [122] research stating that proactivity may not be overloading drivers furthermore points to linguistic complexity being a factor for successful proactive interactions. This paper examines user preferences for the formulation of German proactive prompts in an automotive setting and aims at closing the gap of to date insufficient design guidelines for proactive in-car prompts, by addressing the following research question: *Are there best practices for the formulation of proactive prompts?*

The paper is structured as follows: chapter 2 reviews relevant literature around proactivity and the formulation of in-car prompts. The subsequent method section defines linguistic parameters constituting a) complex and b) suggestive/imposing language. Proactive in-car prompts are then modified regarding these parameters on syntactical, lexical, and grammatical levels. Due to currently insufficient frameworks for measuring the usability of single VA prompts, a Likert scale is developed and validated for the purpose of this study. By means of a driving simulator study, proactive prompts are evaluated in a concrete driving scenario to identify linguistically driven best practices for proactive in-car features. Chapter 4 analyses study participants’ formulation preferences for proactive prompts, which are discussed further in chapter 5. In chapter 6, we conclude that study participants indeed have formulation preferences regarding complexity and suggestive language. With the aid of these preferences, concrete linguistic design guidance for proactive in-car interactions in German is formulated.

2 RELATED WORK

The following chapter is to review relevant literature embedding this research project in a larger context. Essential links are drawn to previous and related research regarding current best practices for proactivity and the formulation of VA prompts.

2.1 Proactive Voice Assistants

Current reactive VAs wait for users to initialize interactions, while proactive agents make a shift to more conversational and natural interaction patterns. Proactive features are thereby found to be popular and among users' envisioned features for perfect voice assistants [121, 141, 148]. With proactivity breaking up currently predominant question-answer patterns with VAs, the feature opens up room for increasingly collaborative and augmented communication [2]. Braun et al. [14] thereby show that users want their proactive VAs to display authentic, and human-like personas. Furthermore, they call for customizable and personalizable proactive agents and functions, which is supported by Koch et al. [69]. Kraus et al. [72] find that systems providing proactive notifications and suggestions lead to increased trust compared to strictly reactive VAs. The researchers add that proactive suggestions can relieve stress in trying situations by supporting decision making processes through confirmations or positive reinforcement. They conclude that especially novice users can profit of the enhanced interaction possibilities of proactive agents [72].

Proactivity has proven to not only be a useful feature in contexts like the smart home [114, 148], but also in in-car settings. In comparison with a reactive assistant, a proactive assistant was found to be similarly demanding and receive similar SASSI ratings in terms of fun and usefulness [122]. Reaction times on the other hand decreased for proactive conditions compared to reactive conditions [122]. In general, driving constitutes a rich environment for proactive suggestions. In the car, proactive use cases related to the driving task include refueling, finding parking spaces, or suggesting faster routes. Furthermore, comfort-related use cases provide intriguing possibilities for proactivity. Multiple studies explored proactive well-being use cases with breathing and mindfulness exercises and found beneficial stress-reducing effects [5, 69, 70, 105].

To carve out a catalogue of appropriate moments for proactive in-car voice interactions, Semmens et al. [127] asked drivers “Is now a good time [to interrupt]?”. While they find a high amount of individual variation (supporting Braun et al.’s [14] call for customizable proactive assistants), their question is answered with “yes” in 77.9% of cases [127]. Still, a note of caution needs to be sound for proactive agents both inside and outside the car. Although the above-mentioned research points to proactivity being a useful and helpful feature, it can also be perceived as an unwelcome intrusion. In studies by Zargham et al. [148] and Reicherts et al. [114], users raise concerns regarding their agency, stating they fear a loss of control or felt patronized. Zargham et al. [148] even speak of a proactivity dilemma and Reicherts et al. add that “proactive VAs need to strike the right balance between being helpful and being intrusive” [114, p. 2]. Hence, timing and external circumstances for proactive suggestions need to be considered carefully before interrupting users who are potentially preoccupied with primary tasks. This is especially true for proactivity in the vehicle, where the primary driving task is not only challenging but also security relevant [120]. In case drivers are presented with a demanding secondary task, they are known to engage in compensatory driving behavior and adapt their speed and steering wheel positions [62, 120]. Alternatively, drivers lower their secondary task engagement or opt to not interact with a secondary task at all [63, 106]. Primary and secondary task are hence mutually dependent, in that a demanding primary task negatively impacts secondary task engagement, while a demanding secondary task can lead to poorer driving performance. Balters et al. [5] and Paredes et al. [105] add to this research. In their studies, they explored proactive well-being features and conclude that in-car interventions need to be “subtle, unobtrusive, and easy to engage and disengage with” [105, p. 4]. Considerations around in-car proactivity are therefore largely centered around users’ “interruptibility” and “opportune moments” for interruptions [25, 59, 106, 109]. In order to be “interruptible”, users’ auditory and verbal channels need to be available, meaning they should not be preoccupied with an ongoing conversation [25, 114]. More factors have proven to be important for a well-rounded user experience for proactive agents though. Reicherts et al. [114] find that group settings are a detrimental factor for proactive interactions. Koch et al. [69] support this finding for an in-car context, where acceptance for proactivity decreased with the presence of further passengers in the vehicle. Furthermore, moments with a generally low workload are found to be so-called “best moments” for proactivity, as shown in a study by Iqbal et al [59]. The researchers state that interruptions in best moments lead to a high degree of social attribution, as well as to a low degree of annoyance and resumption lag [59]. Cha et al. [25] extend these findings and add users' concentration, primary task engagement, urgency, and busyness as factors to be considered when timing proactive suggestions. While the momentane mental workload may not be too high in order not to overload users, Cha et al. [25] find that users occupied with a highly engaging but not challenging task are most susceptible to proactive suggestions. Balters et al. [5] specify such scenarios by means of qualitative data, stating that users can imagine proactivity in traffic scenarios like driving on a highway or stopping at a red light. As for social factors, the presence of another person did not influence interruptibility in Cha et al.'s [25] study, but the researchers find less interruptibility in users who are in a bad mood. Beyond personal contextual factors, the researchers furthermore identify movement related as well as social factors playing into interruptibility. Dynamic activities, the transition between activities, as well as entrance and departure are found to be opportune moments for interruptions. These findings are supported by Pejovic et al. [109] who describe that users who transition between activities are susceptible to interruptions. Koch et al.'s [70] findings are partly even more fine-grained and found times of day and weather details to influence acceptance of proactivity. Rain thereby had a negative impact on acceptance of interventions. On the contrary, evenings and weekends were favorable moments for proactivity.

2.2 The State of Design Guidelines for Voice Assistant Prompts

2.2.1 General Design Guidelines

Design recommendations for VAs date back several decades [39, 102, 108, 129]. While studies and practical experiences by e.g., Pearl [108], Vlahos [140], or Nass and Brave [101] advocate for natural, straightforward, and informal interactions with VAs, they oftentimes lack concrete guidelines VA designers can adhere to. Research around factual prompt design guidelines is especially scarce and only a few studies provide linguistic details for designing (in-car) VA prompts.

Research around driver-centric and natural VAs suggests that prompts should be designed in "short and clear segments" [1, p. 157] and points out that processing capacities for prompts are highly user specific. Nonetheless, Alvarez et al. [1] do not define "short and clear" further. Schmidt et al. [124] who explore and assess proactive use cases for in-car VAs add that users appreciate precisely and thoroughly formulated prompts, but do not go into detail on how to achieve this goal.

Semmens et al.'s [127] research implies that the more natural a proactive interaction, the less distracting it is, but here too, concrete formulation recommendations are missing. Further research on how to interrupt users is provided by Nallapaneni [100], who explores the interaction of proactivity and speaking style and found an emotionally styled text to speech (TTS) voice to be rated more attractive than a compared neutral speaking style. Zargham et al.'s [148] and Hofmann et al.'s [56] research on the other hand is concerned with how to interrupt users in terms of how to prepare them for an upcoming proactive suggestion. Both propose the usage of sounds or earcons to get users' attention. In a driving simulator study, Hofmann et al. [56] find that earcons are perceived to be the least distractive way of announcing proactivity, while announcements via speech are ranked highest in terms of usability. Zargham et al. [148] furthermore argue to give users the possibility to verbally confirm or deny proactive interruptions after hearing a respective sound sign. They go on to suggest that a proactive prompt "should be phrased so that it is polite, not imposing, and does not create a feeling of unease, while at the same time being goal-oriented and concise" [148, p. 10]. The researchers thereby mirror above-mentioned design recommendations, but do not provide further support as to how a prompt can be formulated in the suggested manner.

2.2.2 Linguistic Design Guidelines

While research in the previous chapter touches on design guidelines for VA prompts, Meck et al. [94] and Stier et al. [134] explore linguistic-driven design guidelines more thoroughly. The researchers' studies find particular and fine-grained formulation preferences regarding syntax, grammar, and lexis of VA prompts. Stier et al. [134], who compared different syntactical structures in in-car prompts, found syntactically simpler and linear main clauses to be preferred over more nested and complex relative clauses. The researchers partly link these preferences to demographic characteristics such as age and personality traits. Meck et al. extend these considerations around formulation preferences by examining further syntactical parameters like e.g., sentence length and moreover compare grammatical, and lexical parameters [94]. In their exploratory study, the researchers want to determine the extent to which linguistic considerations play a role in the perception of in-car VAs. In total, they analyze 28 linguistic parameters for different types of conversations with VAs in the vehicle: functional prompts (VA responses after being asked to carry out a function), informational prompts (VA responses after being asked for information), and chit chat prompts (VA responses relating to small talk). As in Stier et al.'s [132] study, formulation preferences partly depended on the type of conversation, which Meck et al. [94] link to varying user needs underlying different conversational contexts. Exemplarily: in their study, study participants appreciated the use of filler words for chit chat use cases while they strongly opposed them for informational prompts. The researchers explain that user needs in chit chat conversations are rather informal, natural, and conversational which allows for light elements like filler words. The user need in informational prompts on the other hand is more rigid and information centered. Non-informational filler words become unnecessary elements which only inflate a prompt without adding important information [94]. Next to concrete best practices around prompt formulations for different types of conversations, the researchers conclude their research by defining three user needs underlying conversations with VAs: a suitable level of (in)formality, (im)mediacy, and complexity. (in)formality can thereby be reached by e.g., using an active voice and politeness markers. Complexity is constituted on a syntactical level and is e.g., dependent on sentence structure and sentence length. Lastly, (im)mediacy can be varied by grammatical parameters such as tense and lexical parameters such as filler words and form of address.

As Meck et al. [94] describe the currently most exhaustive set of linguistic parameters with potential formulation preferences, their work is consulted for the selection of study parameters for the present paper. Although hands-on linguistic design guidelines are compiled, previous studies were not tailored to proactive interactions. As the researchers do find an influence of the type of conversation though, formulation preferences established so far might differ between proactive and previously studied conversational contexts. Furthermore, Meck et al. conducted their experiment in form of an online A/B study. Hence, the present study will build on the researchers' study but extend it by two components:

a) a proactive conversational context

b) an in-car study set-up in form of a driving simulator study.

2.2.3 Selection and Description of Linguistic Parameters

Given already existing design guidelines for proactive (in-car) interactions, language complexity [132, 134] as well as suggestive language [105, 114, 148] are identified as essential factors for designing proactive in-car prompts. While linguistic complexity develops on a syntactical level [94, 132], we believe suggestive language to be dependent on grammatical and lexical parameters.

To control the degree of prompt complexity, the syntactical parameters sentence structure, sentence length, as well as the position of sub-clauses can be modified. Loss of agency emerged as a substantial user concern in both Reicherts et al. [114] and Zargham et al.'s [148] studies. To address this concern, the parameter form of address (the form of reference a VA employs to address users) is chosen from Meck et al.'s [94] parameter pool. Lastly, to refrain from imposing language, the parameters voice (active vs passive voice) and politeness are adopted from Meck et al. [94]. The following paragraphs explain all selected study parameters and their explored manifestations further. An overview over all study prompts can be found in Appendix A.

Sentence Structure. Manifestations: *parataxes, hypotaxes, multi-clause sentences*. Intricate sentence structures such as multi-clause sentences (MCS), which consist of main clauses with two or more nested sub-clauses, are known to be complex to process for listeners. In comparison, more straightforward parataxes and hypotaxes consume less processing capacities [26, 131]. Paratactical sentences, meaning sentences with sequential main clauses, consist of separated and distinct processing units, which facilitate language understanding. Regarding their processing capacities, hypotactical sentences can be placed between parataxes and MCS. Previous studies have painted a mixed picture regarding preferences for sentence structures. While study participants preferred parataxes in Stier et al.'s [132] study, Meck et al. [94] found a preference for hypotactical sentences.

Sentence Length. Manifestations: *short, medium, long*. The fleeting nature of speech requires high attentional and memory skills to build a situational model of comprehension [65, 145]. In general, processing demands increase with prompt length, but auditory information processing is highly user specific and depends on an individual's surroundings [1]. Furthermore, prompts need to strike the right balance between being short and concise, but still meeting users' information needs and provide them with all necessary facts. A previous study found information needs to differ between conversational contexts, such as functional prompts vs informational prompts [94]. While short prompts were preferred in functional prompts, study participants preferred long prompts in informational prompts.

Position of Sub-Clauses. Manifestations: *prepositive, postpositive*. The position of a sub-clause is determined in relation to its concomitant main clause and can either precede (prepositive sub-clause) or follow it (postpositive sub-clause). Sub-clauses describe a main clause further and their position is important in terms of information packaging and information processing. In conversations, interlocutors will package information in a manner that is most appropriate for the context and most easily processible by their conversational partner [26]. Main clauses thereby contain the so-called theme, the topic of a conversation, while sub-clauses contain the rheme, meaning they describe the theme further [128]. Sub-clauses are introduced with conditional, temporal, or causal conjunctions, which contain crucial information regarding what, when, why, or how something is happening in the main clause. While prepositive sub-clauses were preferred in functional prompts, postpositive sub-clauses reached higher ratings for informational and chit chat prompts.

Form of Address. Manifestations: *1st person singular, 2nd person singular, 1st person plural*. Reactive VAs mostly function as service-oriented assistants and generally address users with "I", as in "I can help you with …". This form of self-referencing emphasizes how a VA can assist a given user. Limerick et al. [83] highlight a diminished sense of agency going along with using voice as an interaction modality compared to touch though. Proactivity enhances this feeling as conversations are not triggered from the users' side, but are carried out unsolicitedly by the VA [114, 148]. Addressing users with "you" as in "Do you want to …" allows a linguistic hand-over of control from the VA towards the user. Other forms of address linguistically hand over agency from VAs to users, by addressing them with "you", as in "You can activate…". In rare cases, joint referencing in form of "we" – forming a unit between VA and user – is adopted. According to Watzlawick [142], communication can either be complementary or symmetrical, which can be reflected through the form of address. In their study, Meck et al. [94] found a preference for joint referencing in chit chat and informational prompts, whereas the usage of service-oriented referencing in form of "I" was preferred in functional use cases.

Politeness. Manifestations: *politeness, no politeness*. Politeness is an inherently social trait of human language and an intuitive design choice for service-driven VAs. Still, politeness is a debated concept in HCI, as it stretches "social boundaries" [32, p. 1], potentially leading to Uncanny Valley effects. Politeness can be expressed grammatically (e.g., "May I ask you to repeat that?") as well as through lexical politeness markers (e.g., "Can you please repeat that"). Meck et al. [94] found grammatical politeness to be rated poorly by study participants. Preferences for lexical politeness differed between prompt types. While half of all study participants preferred lexical politeness in functional prompts, informational and chit chat prompts were preferred without politeness markers [94].

Voice. Manifestations: *active*, *passive*. The selection of an active or a passive voice influences the emphasis in a given sentence. In actively formulated prompts, the emphasis is on the agent itself, while a passive voice shifts emphasis from an agent to a proposed action [26]. The active voice hence focuses on who carries out a given task. On the contrary, the passive voice focuses on the task itself. Furthermore, an active voice is more conversational, while a passive voice can be found in written form and official language [131]. While active prompts are preferred in service-driven and function-oriented interactions, a higher tolerance for passive prompts was found for information-driven conversations [94].

Table 1 summarizes the linguistic parameters outlined above, including their potential manifestations. As discussed, syntactical parameters contribute to linguistic complexity, while lexical and grammatical parameters compose the tone of voice.

Table 1: Overview and Clustering of Study Parameters

Parameter	**Complexity**		**Parameter**	**Suggestive Language**	
	Simple	**Complex**		**Suggestive**	**Imposing**
Sentence Structure	parataxes	hypotaxes, MCS	*Form of Address*	2nd person singular	1st person singular, 1st person plural
Sentence Length	short	medium, long	*Voice*	passive voice	active voice
Position of Sub-Clauses	prepositive	postpositive	*Politeness*	politeness	no politeness

3 METHOD

To answer the research question outlined in the introduction, linguistic parameters deemed important for proactive interactions were extracted from related research. Due to currently insufficient frameworks for measuring the usability of single VA prompts, a Likert scale is developed and validated for the purpose of this study. Subsequently, a within-subjects driving simulator study is conducted to evaluate best practices for the formulation of proactive prompts in German.

3.1 Development of a Tool for Measuring Formulation Preferences

3.1.1 Development of the Study Scale

A large body of research can be drawn upon to measure the overall system usability of VAs [57, 67, 76]. However, no standardized and validated state-of-the-art questionnaire exists to identify usability on a prompt level. While Meck et al. [94] compared best practices for the design of prompts via A/B studies, Stier et al. [132] fell back on Likert scales. A/B studies are a proven approach to compare two or more factors, such as prompts, but due to the fleeting nature of speech, A/B studies are not necessarily appropriate for comparing auditory stimuli. Furthermore, the present study takes place in a driving simulator and interactions with the VA are presented as a secondary task, making the A/B comparison of prompts, which partly only differ in nuances, unfeasible. Hence, formulation preferences are queried with the help of Likert scales which are intuitive and conveniently answerable while driving [132]. In line with recommendations from Laugwitz et al. [76] and Klein et al. [67], the Likert scale developed for the present study comprises four items measured on a seven-point scale. Drawing from Stier et al. [132] and the UEQ+ [67], *comprehensibility* as well as *naturalness* are adapted as items. Comprehensibility thereby not only serves as an item, but additionally allows the comparison of study prompts' intelligibility. To complete the scale, the UEQ+, which is specifically tailored to assess the usability of VAs, is consulted. The UEQ+ contains the scales response behavior, response quality, as well as comprehensibility [67]. As the present study aims at examining best practices for the formulation of proactive prompts, the UEQ+ scale response quality is consulted. Response quality contains the items suitability, usefulness, helpfulness, as well as intelligence [67]. The last item, *intelligence*, is adapted as is and added to the Likert scale. Suitability, usefulness, and helpfulness on the other hand describe a prompt's contents rather than its formulation and are hence excluded. Due to the current lack of validated scales for assessing the usability of individual prompts, it is unclear how fine-grained users can evaluate prompts. Hence, the general item *positivity* is chosen as fourth and final study item. The final study scale can be seen in Figure 1. *Naturalness*, *intelligence*, and *positivity* are adapted verbatim to the study scale. *Comprehensibility* was contrasted on a continuum from *very simple* to *very difficult*.

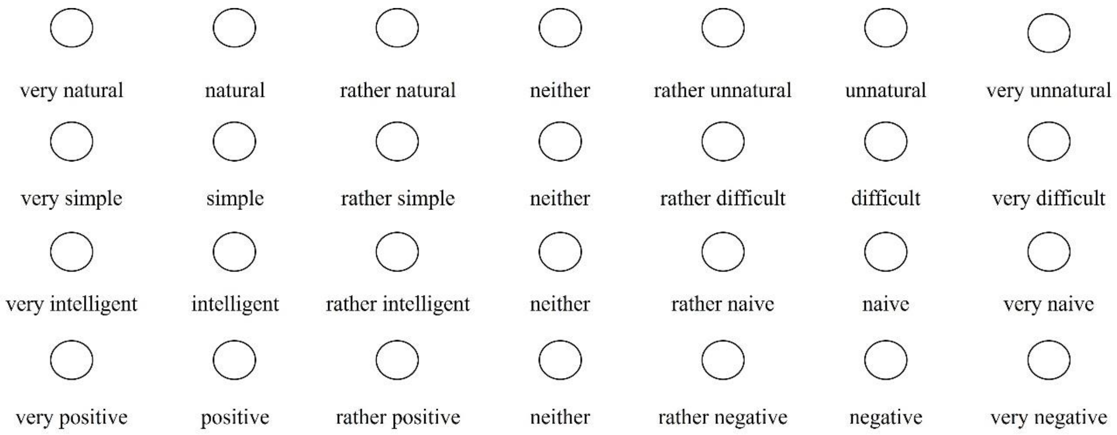


Figure 1: Likert Scale used in the Driving Simulator Study

3.1.2 Reliability Testing of the Study Scale

To validate the developed study scale, a preliminary crowdsourcing study was conducted online. 150 study participants were invited to take part in the experiment. They were presented with the same 15 prompts that were planned to be used in the driving simulator study. Their task was to rate these 15 prompts on the newly developed Likert scale proposed in the previous chapter (see Figure 1). As such, 60 prompt assessments were obtained per study participant which yielded 9000 prompt assessments in total (150 study participants x 15 prompts x 4 ratings regarding naturalness, comprehensibility, intelligence, and positivity).

These assessments were used to measure the scale's internal validity. First, Cronbach's alpha was computed and splithalf reliability tests were conducted. The study scale was found to be highly reliable (4 items; α=.9) [136]. Secondly, the internal consistency was calculated my means of a splithalf approach with 5000 random splits revealed a Guttman lambda 2 splithalf internal consistency of λ-2=.9, 95%CI [.87, .92], which is interpreted as sufficient [22]. Due to the obtained measures described above, the proposed scale was adopted as study scale for the driving simulator study.

3.2 Hypotheses

To answer this paper's research question, linguistic parameters forming complex as well as suggestive language were discussed in chapter 2.2.3. Considering related research, the following hypotheses emerge:

H1: Less complex language is preferred over complex language in proactive in-car interactions.

H2: Suggestive language is preferred over imposing language in proactive in-car interactions.

Linguistic complexity manifests on syntactical levels and is dependent on sentence structure, sentence length, and the position of sub-clauses. To reduce a prompt's complexity, parataxes are hypothesized to be the preferred sentence structure in proactive scenarios. Furthermore, we hypothesize that short prompts are preferred by study participants. By means of conjunctions, prepositive sub-clauses directly indicate the reason for a proactive interruption. Hence, it is hypothesized that prepositive sub-clauses are preferred in proactive contexts. Our hypotheses are:

H1.1: paratactical sentences are rated most comprehensible, intelligent, natural, and positive.

H1.2: short sentences are rated most comprehensible, intelligent, natural, and positive.

H1.3: prepositive sub-clauses are rated most comprehensible, intelligent, natural, and positive.

Suggestive rather than imposing language manifests on grammatical as well as lexical levels. In passively formulated prompts, a proactive agent can direct the focus to a proposed function rather than to itself. As such, we hypothesize passive prompts to be preferred over active prompts. Language including politeness markers constitutes a less imposing and rather suggestive speaking style. Hence, a preference for polite prompts is hypothesized. To counteract users' feeling of diminished agency, the form of address can act as a linguistic handover of control. Hence, we hypothesize that referencing users with "you" is preferred by study participants.

H2.1: 2nd person singular is rated most comprehensible, intelligent, natural, and positive.

H2.2: passive voice is rated most comprehensible, intelligent, natural, and positive.

H2.3: politeness is rated most comprehensible, intelligent, natural, and positive.

3.3 Use Cases and Study Prompts

Six proactive use cases were realized for the driving simulator study: 1) availability of a faster route, 2) parking suggestions, 3) intelligent destination proposals, 4) customizing the navigation map, 5) offering to activate a relaxing mode, and 6) information on the remaining fuel range.

Comparison prompts were modified regarding the linguistic parameters described in chapter 2.2.3 and differed in only one parameter at a time. In a concrete example: the parameter *voice* has two manifestations, namely *active voice* and *passive voice*. Two comparison prompts are formulated, expressing both manifestations respectively, as can be seen in Figure 2.

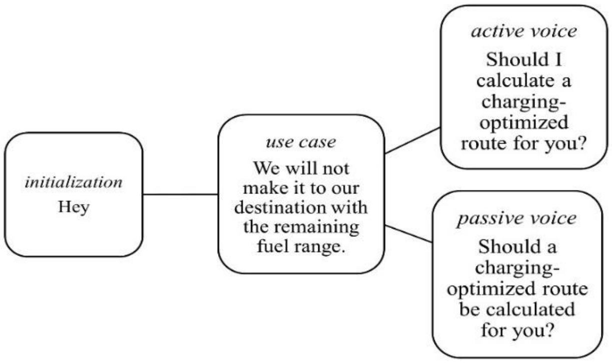

initialization
Hey
use case
We will not make it to our destination with the remaining fuel range.
active voice
Should I calculate a charging-optimized route for you?
passive voice
Should a charging-optimized route be calculated for you?

Figure 2: Example of Comparison Prompts for the Parameter voice

All prompts underwent measures to ensure their suitability as study prompts as well as their comparability as comparison prompts. In a first step, complexity of prompts was measured by means of the so-called LIX[17] readability index [82]. LIX' complexity calculation is based on the overall number of words and clauses, the average clause length in prompts with more than one sentence, as well as the number of long words with more than six characters. The LIX ranges from 20 (very low) to 60 (very high). To ensure comprehensibility and intelligibility of prompts, prompts with a high or very high LIX index of above 55 were eliminated from the study. Except for sentence structure and sentence length, where varying complexity arises specifically because of syntactical differences, comparison prompts only qualified as such if they reached the same degree of complexity. Furthermore, varied parameters (e.g., voice) were positioned in prompts' first or last sentence to make use of primacy and recency effects [37]. All interactions followed the same pattern, shown in Figure 3 below.

[17] LIX is the acronym for Swedish **L**äsbarhets**i**nde**x**, i.e., readability index in English.

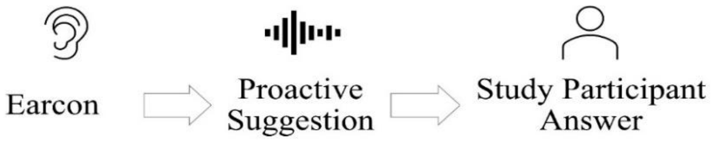


Figure 3: Structure of Study Interactions

As proposed by Reicherts et al. [114] and Hofmann et al. [56], an earcon signaled the beginning of a proactive interaction, which was followed by the proactive suggestion itself. In total, each study participant rated 15 prompts whose distribution is shown in Table 2 below.

Table 2: Overview over Prompt Distribution

Syntactical Parameters: **8 Syntactical Prompts**		**Lexical Parameters:** **5 Lexical Prompts**		**Grammatical Parameters:** **2 Grammatical Prompts**	
Sentence Structure	Hypo Para MCS	*Form of Address*	"I" "you" "we"	*Voice*	Active Passive
Sentence Length	Short Medium Long	*Politeness*	Yes No		
Position of Sub-Clauses	Prepositive Postpositive				

3.4 Driving Simulator Study

3.4.1 Design and Conduct of the Driving Simulator Study

To investigate possible prompt preferences for proactive prompts while driving, a within-subjects driving simulator study was realized. Prompts were presented to subjects in randomized order to counteract sequence effects. Table 3 provides an overview over study variables and factor levels.

Table 3: Overview over Dependent and Independent Variables

Factor	**Factor Levels**			**Dependent Variable**
Sentence Structure	parataxes	hypotaxes	MCS	Likert ratings for naturalness, comprehensibility, intelligence, and positivity
Sentence Length	short	medium	long	
Form of Address	"I"	"you"	"we"	
Position of Sub-Clauses	prepositive	postpositive		
Politeness	politeness	no politeness		
Voice	active	passive		

A stationary vehicle mock-up surrounded by a 180° screen served as environment for the driving simulator study as can be seen in Figure 4.

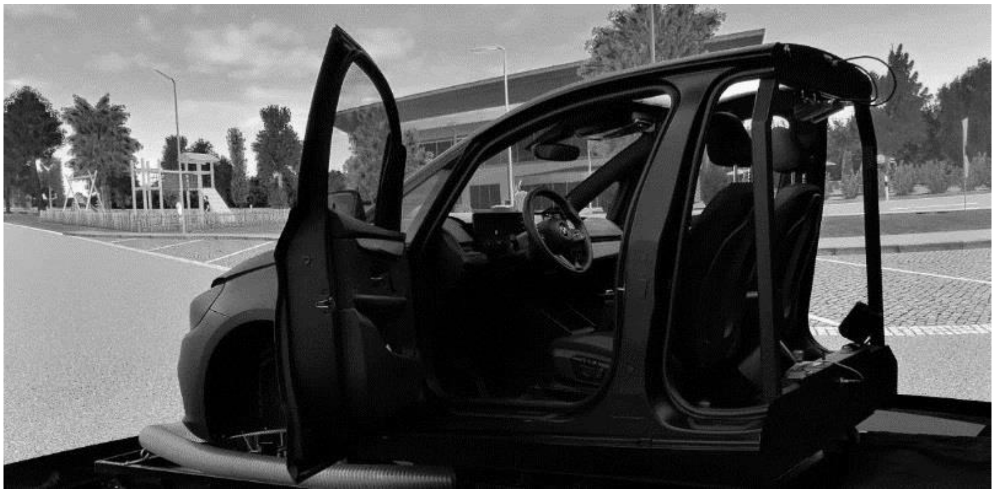


Figure 4: Driving Simulator Study Setup

The vehicle mock-up was equipped with a VA which could be administered from the operator desk in a Wizard of Oz manner. To account for the mutual dependency between primary and secondary task, the driving task was designed simple and straightforward. Driving simulator studies continually report low secondary task engagement and resulting poor user evaluations due to primary task overload [62, 63, 106, 120]. Furthermore, extensive research has shown that proactivity is highly context-dependent and unsuitable when users and drivers are concerned with demanding primary tasks [122, 127]. Resultingly, a highway setting with low traffic density was chosen as driving scenario to allow study participants to focus on the prompt evaluation task. The primary driving task consisted of following a lead vehicle on the right highway lane with a speed of 100km/h. Conditions did not vary between study participants and throughout the experiment drive. Due to the low traffic density setting and the straightforward and simple driving task, measures around driving performance were not collected.

Upon arrival, subjects gave informed consent and filled out a questionnaire with demographic questions. The experimenter then led them to the vehicle mock-up and familiarized them with the experiment procedure as well as the Likert scales to be used for the evaluation task. A copy of the scales remained in the mock-up throughout the experiment. Subjects were asked to rate 15 proactive prompts on four seven-level Likert scales regarding their overall positivity, intelligence, comprehensibility, and naturalness. After subjects were seated in the vehicle mock-up, the experimenter retreated to the operator desk and verified the audio connection to the mock-up. Subsequently, subjects were asked to start a three-minute familiarization drive to get accustomed to the vehicle and the route. After completing the familiarization, subjects received a test prompt and were asked to rate it on the Likert scales before the actual experiment started. The prompt evaluation task then developed according to Figure 3 (see chapter 3.3). Participants were asked to answer to suggestions naturally and as if they were proposed to them during a regular drive. After hearing each prompt, they rated the respective prompt on the Likert scales. As the goal of the present study was to find potential preferences regarding concrete prompt formulations, subjects were requested to focus on the formulation of a prompt and neglect the TTS voice presenting it. After the rating of a prompt was completed, the experimenter waited for a full minute before triggering the next interaction. In total, the drive lasted approximately 40 minutes per participant. After completing all evaluations, subjects were being seen out by the experimenter.

3.4.2 Subjects

The sample size of the study was determined a priori through a power analysis in G*Power for a repeated measures ANOVA [43]. The effect size was estimated according to Stier et al. [132], who found effect sizes from r=.23 to r=.32. Conservatively, the effect size for the present study was set to r=.23. With α=.05 and β=.95, the needed sample amounted to 48 participants. Ultimately, N=60 participants were invited to take part in the study, to allow for a buffer in case of e.g., technical problems. n=2 participants had to terminate their attendance due to motion sickness during the familiarization drive and were hence excluded from the study. 59% of subjects identified as male, 41% as female. Age of participants ranged from 20 to 59 with a mean of 34.46 years. The age group of 18-35 years made up 53% of participants, and accordingly, study participants from age 35 to 60 made up 47% of subjects. Asked about their usage of in-car VAs, most participants set their usage to several times a week (33%), followed by several times a month (21%). Anecdotally, the experimenter did not report notable differences in study participants' response times when accepting/denying proactive suggestions. Study participants did not indicate problems with the driving task (such as e.g., adhering to the speed limit or keeping the lane), nor did the experimenter report deviating driving behavior.

4 RESULTS

N=58 subjects rated 15 prompts on four seven-level Likert scales regarding their overall positivity, intelligence, comprehensibility, and naturalness.

4.1 Statistical Analysis

In total, 77% of all proactive suggestions were answered with “yes” by study participants. Separated per use case, the approval rate can be seen in Table 4.

Table 4: Approval Rate of Proactive Use Cases

Use Case	Approval	Use Case	Approval
availability of a faster route	83%	*customizing the navigation map*	67%
parking suggestions	87%	*offering to activate a relaxing mode*	61%
intelligent destination proposals	83%	*information on the remaining fuel range*	96%

Prompt preferences were queried on Likert scales and the obtained data was found to not be normally distributed. Due to the data structure, non-parametric Cumulative Link Mixed Models using the *clmm()* function in the ordinal package in R [116] were fitted for the statistical analysis. The dependent variable *prompt evaluation* was predicted as a function of the independent variable *linguistic parameter* interchanged between two prompts. Demographic factors age and gender, as well as experience with in-car VAs served as fixed factors. Due to the repeated-measures design of the study, subjects were introduced as random factors. The model translates to:

clmm(prompt evaluation ~ linguistic parameter + age + gender + experience + (1|study participants)

All results are reported for α=.05. Overall, the Likert item comprehensibility did not produce significant results across parameters. Intelligence on the other hand yielded significant results for all parameters except voice. Naturalness was a significant factor in ratings for politeness, while form of address was rated significantly different in terms of overall positivity.

In terms of intelligence, parataxes were subjects’ preferred sentence structure compared to hypotactical sentences. For sentence length, short prompts with 20 words were evaluated as significantly more intelligent than long prompts (30 words). For form of address, “I” received better positivity ratings than “you”. Regarding intelligence, “you” was rated significantly more intelligent than “I” though. Concerning position of sub-clauses, prepositive sub-clauses were preferred over postpositive sub-clauses and deemed significantly more intelligent. Significant differences were found for politeness for the Likert dimensions intelligence and naturalness. Polite prompts were evaluated significantly more intelligent than their less polite counterparts, while less polite prompts were rated as significantly more natural than polite prompts. As hypothesized for voice, passive voice was preferred over active voice by subjects, although results were not significant. A complete overview over Likert ratings, including means and standard deviations can be found in Appendix B. Appendix C comprises an exhaustive overview over all tested results. Table 5 provides an overview over statistically significant parameters.

Table 5: Overview over Statistical Results

Parameter	**Statistical Results for Intelligence**
Sentence Structure: hypo/mcs	R^2=0.56 (marg R^2=0.04), OR 0.71, 95% CI [0.71-0.72], p<.001
Sentence Structure: hypo/para	R^2=0.56 (marg R^2=0.04), OR 0.50, 95% CI [0.50-0.50], p<.001
Sentence Length: long/med	R^2=0.62 (marg R^2=0.07), OR 1.84, 95% CI [1.83-1.84], p<.001
Sentence Length: long/short	R^2=0.62 (marg R^2=0.07), OR 0.98, 95% CI [0.97-0.98], p<.001
Form of Address: I/we	R^2=0.76 (marg R^2=0.01), OR 0.67, 95% CI [0.66-0.67], p<.001
Form of Address: I/you	R^2=0.76 (marg R^2=0.01), OR 1.54, 95% CI [1.53-1.55], p<.001
Position of Sub-Clauses	R^2=0.82 (marg R^2=0.01), OR 0.47, 95% CI [0.47-0.47], p<.001
Politeness	R^2=0.70 (marg R^2=0.16), OR 0.71, 95% CI [0.71-0.71], p<.001
Mood	ns
Parameter	**Statistical Results for Naturalness**
Politeness	R^2=0.66 (marg R^2=0.08), OR 0.92, 95% CI [0.92-0.93], p<.001
All other parameters	ns
Parameter	**Statistical Results for Positivity**
Form of Address: I/we	R^2=0.67 (marg R^2=0.14), OR 1.65, 95% CI [1.65-1.66], p<.001
Form of Address: I/you	R^2=0.67 (marg R^2=0.14), OR 1.44, 95% CI [1.43-1.45], p<.001
All other parameters	ns
Parameter	**Statistical Results for Comprehensibility**
All parameters	ns

4.2 Demographic Analysis

As explained in the previous chapter, Cumulative Link Mixed Models were calculated to account for the ordinal nature of the data [116]. Age was found to have an influence on subjects’ evaluations regarding positivity and intelligence of sentence length and form of address. While younger subjects found short prompts to be more intelligent than long prompts, older participants showed the opposite tendency and rated long prompts as more intelligent: $R^2=0.62$ (marginal $R^2=0.07$), OR 1.08, 95% CI [1.01-1.16], p=.03. Where younger subjects rated “you” as overall more positive, older subjects found “I” more positive: $R^2=0.67$ (marginal $R^2=0.14$), OR 1.10, 95% CI [1.02-1.19], p=.01. No significant influence was found for gender and previous experience with VAs.

4.3 Qualitative Insights

A semi-structured interview was conducted after the simulator study. It inquired whether study participants noticed differences in the proactive prompts they experienced. Furthermore, they were asked to describe these differences if possible. Lastly, they were encouraged to leave general feedback around proactivity and the experienced use cases.

58.62% of study participants specified to have noticed differences in the presented prompts. Another 8.62% partly observed differences, while 32.76% indicated to not have noticed differences at all. Asked about the nature of differences, 79.49% attributed contrasts in prompts to lexis, 33.33% to syntax, and 7.69% to the TTS voice presenting the prompts. Differences in syntax were connected to prompt length (53.85%), complexity (30.77%), as well as sentence structure (30.77%).

General feedback around proactivity and the experienced use cases was transcribed and clustered into four themes: *content*, *wording*, *suitability*, and *context*.

Content. 50 study participants shared their impressions regarding the content of proactive interactions. 18 participants stated that prompts lacked detailed information and explanations necessary for understanding the proposed functions in their entirety. On the contrary, 32 participants criticized study prompts for being too long and complex. They expressed their wish for short prompts, comprising only necessary information without describing the reason behind proactive interruptions in detail.

Wording. Wording-related remarks were given by 40 study participants. Participants expected proactive proposals to be presented in natural and informal language. Comments on wording were partly highly specific and criticized or commended specific words. The verbal proactive introduction was varied between “Hey” and “Just for your information”. Study participants were divided regarding the perceived suitability of the introductions and asked for varying introductory sentences. While some participants found “Hey” to be too informal and intrusive, “Just for your information” was deemed too formal.

Suitability. 21 study participants disclosed concerns around the suitability of proactive suggestions. They stated that suggestions need to be useful and helpful. Furthermore, they should be relevant to the concrete driving situation and propose directly applicable functions. Suggestions should add general and informational value, by e.g., pointing out prospective problems and providing anticipatory solutions. Some participants shared their preference for switching from speech to touch to confirm or deny proactive suggestions or receive further information.

Context. 11 study participants mentioned context-related concerns, such as fear of being overloaded or disturbed in conversations with co-drivers. Furthermore, they indicated that a proactive agent should not repetitively propose use cases and rather memorize whether a proposal was accepted or denied in previous interactions. Participants emphasized the importance of usefulness and motivation behind proactive suggestions as well as an existing trust relation between themselves and their agent.

5 DISCUSSION

While previous studies have focused on interruptibility in terms of when to interrupt users, this study is concerned with interruptibility in terms of how to interrupt them. In this paper, findings are discussed to shed light on concrete linguistic guidelines for formulating proactive VA prompts under consideration of demographic factors. In line with existing research, more than two-thirds of proactive suggestions in the driving simulator study were accepted, establishing proactivity once more as an important and sought-after feature. Use cases with relevance to the driving task were thereby accepted more frequently than non-driving or comfort-related use cases. While proactive information on the remaining fuel range was accepted in 96% of cases, offering a relaxing mode or customizing map settings received 30% less approval. These findings mirror related research from Schmidt et al. [122] and show that proactive suggestions are most suitable if they context-sensitively relate to the ongoing primary (driving) task. The finding can also be interpreted in light of Reicherts et al.’s statement, saying that “proactive VAs need to strike the right balance between being helpful and being intrusive” [114, p. 2]. The more helpful and relevant a proactive suggestion is in a given situation, the more likely it is accepted. Related to this conclusion, approval ratings of prompts were not found to vary notably regarding linguistic parameters. This finding underlines that proactive suggestions are approved depending on their content. However, while linguistic parameters do not seem to influence the general approval of proactive suggestions, they do play a role in the perception of prompts as e.g., intelligent.

5.1 Discussion of the Study Scale

Overarchingly, linguistic parameters did not differ significantly in their evaluations on the Likert dimension comprehensibility. Naturalness was only a significant factor in subjects' evaluations of politeness. While positivity only proved to be a significant influence factor for form of address, all parameters except voice were rated significantly different regarding the perceived intelligence of formulation options. The present study hence found linguistic preferences for the formulation of proactive prompts. Still, it needs mentioning that these preferences were not significant for all queried Likert scale dimensions. Intelligence was found to be the only Likert dimension where prompt evaluations differed significantly between all parameters, except voice. Proactivity is a conversational and human-like speech pattern. Following Norman's gulfs of execution and evaluation [103], the more human-computer interactions resemble human-human interactions, the more intelligence is expected of them. Various studies [34, 86] have examined this theorem and found humanness in VAs to "spark comparisons with human assistants" [34, p. 1], which is an explanation for the significant results on the Likert dimension intelligence. A possible explanation for prompts' equal ratings regarding comprehensibility lies in their standardization using the LIX readability index [82]. Study prompts were checked for their complexity using the LIX and only qualified as comparison prompts if they did not differ regarding their intricacy. While this step is important to rule out differences in evaluations due to differing information processing demands when processing study prompts, it could have been a hindering factor when comparing prompts' comprehensibility. As for naturalness not reaching significantly different evaluations, subjects were instructed to rate the formulation of a prompt rather than the TTS voice presenting it. It is possible that naturalness was not a tangible concept for study participants on the formulation level of a prompt, although Stier et al. [132] found significantly different naturalness ratings for different syntactical structures. Lastly, positivity is potentially too broad a concept to depict changes between prompts in a significant manner. Although the designed Likert scale was able to reach very good values in reliability tests as well as reveal significant differences in the evaluation of prompts, it does not yet seem to be the optimal measurement instrument for testing the usability of individual prompts.

5.2 Discussion of Linguistic Preferences

Regarding prompt complexity, which manifests on a syntactical level, study participants found paratactical as well as short prompts to be the most intelligent syntactical structure for proactive prompts. Paratactical sentences structure information in small and distinct processing units. Especially while driving and for unsolicited proactive interactions, straightforward parataxes aid users in efficiently processing a prompt. This result is contrary to findings from both Meck et al. [94] and Stier et al. [132], who found a preference for more complex hypotactical sentences. However, these studies were not looking into proactive use cases. The need for simple and unobtrusive language which increases in proactive scenarios is best catered to by a paratactic sentence structure. The same holds true for short sentences. Study participants preferred prompts with a word count of 20 words over longer prompts and found them more intelligent. Again, as found for less complex sentence structures, more concise prompts are less complex to process. With this finding, our research is in line with related work [94, 132]. Results showed prepositive subordinate clauses to be the preferred formulation option regarding position of sub-clauses. By means of the conditional, temporal, or causal conjunctions introducing them, prepositively put sub-clauses directly indicate the reason behind a proactive interruption. This is especially urgent in an in-car environment where proactivity is potentially security-relevant [122, 127], making reasonable information packaging and quick information processing all the more important. As such, significant preferences regarding complex language were found, which means that H1 – Less complex language is preferred over complex language in proactive in-car interactions – with H1.1-H1.3 can be accepted with the restraints discussed in chapter 5.1.

Regarding form of address, “I” was rated significantly more positive than “you” and “we”. In turn, “you” was rated significantly more intelligent than “I” and “we”. In line with related research [114], the positive ratings for “I” indicate that VAs are primarily seen as service-oriented assistants, carrying out functions for users. Especially in proactive contexts, where a VA is unsolicitedly addressing users, this service-orientation is best addressed by focusing on what the assistant can do for the user. On the other hand, and contrary to this finding, “you” was evaluated to be more intelligent than “I”. This can be explained by falling back on research around voice and agency by Limerick et al. [83], who found a diminished sense of agency for speech compared to touch. Referencing users with “you” allows a linguistic hand-over of control from the VA to the user, putting users’ agency in the foreground rather than focusing on what a VA can do for them. This counteracts the feeling of reduced control and explains the results for intelligence ratings. In sum, while the service oriented “I” is rated more positive, “you” is perceived as the more intelligent formulation option. Although previous research has shown an increased feeling of collaboration in proactive interactions [2], joint referencing of the VA and the user with “we” was rejected by study participants.

Results for voice did not reach critical levels of significance. While the emphasis in an actively formulated prompt lies on the agent itself, emphasis shifts towards the proposed action in a passively formulated prompt. Although a previous study found best practices for voice [94], these could not be replicated for proactive prompts. Lastly, the use of politeness was preferred over less polite prompts and polite prompts were rated significantly more intelligent than their less polite counterparts. Interestingly though, less polite prompts were rated significantly more natural than polite prompts. With politeness being an inherently social trait of human language, it stretches social boundaries, as proposed by Clark [32]. Hence, study participants potentially felt it was more natural (as in appropriate) for a VA to not be polite. At the same time, a polite prompt could have been rated as more intelligent, as politeness is a sociolinguistic construct of human speech. Lee et al. [78] second this, as they found drivers to evaluate polite autonomous vehicles to be more sociable and trustworthy. Furthermore, politeness strategies increased collaboration between the driver and the vehicle in their studies. As these results are only partly consistent with our findings though, we believe cultural aspects to play a role in linguistic preferences for prompts. As such, H2 – Suggestive language is preferred over imposing language in proactive in-car interactions – with H2.1-H2.3 can only partly be accepted.

While concrete formulation preferences for proactive in-car interactions can be obtained, these preferences are not as resounding as for other types of conversations with VAs, where a multitude of best practices was found for speaking style [94]. The present study was conducted in German and results are likely language- and culture-dependent. Our results deliver strong, consistent support for the significance of linguistic complexity for proactive prompt preferences. Furthermore, a suggestive rather than imposing speaking style emerged as an important factor to consider when designing proactive in-car interactions. However, results around speaking style paint a comparably more mixed picture and recommendations for speaking style are not as explicit and consistent as for linguistic complexity. These divergent results for speaking style indicate that preferences for speaking style are a) highly subjective and b) most likely language- and culture-dependent.

Table 6 summarizes our findings and provides an overview over formulation preferences and hence best practices for formulating proactive prompts:

Table 6: Overview over Formulation Preferences for Proactive Prompts

Parameter	**Best Practice**	**Parameter**	**Best Practice**
Sentence Structure	parataxes	*Form of Address*	"I": positive | "you": intelligent
Sentence Length	short (~20 words)	*Politeness*	no politeness: natural | politeness: intelligent
Position of Sub-clauses	prepositive	*Voice*	no significant preferences

In our study, language complexity emerged as a crucial factor for proactivity. Suggestive language on the other hand is much less tangible, rather subjective, and likely language- and culture-dependent. We agree with the large body of research that deems suggestive and unobtrusive language important for proactive in-car interactions [114, 123, 127, 148]. However, we wish to add further points to this design recommendation: subjectivity and language and culture dependency. This addition is necessary as speaking style cannot be formalized and generalized as well as language complexity. Furthermore, we believe that our results support that *when* to interrupt drivers still seems to play the most important role regarding acceptance of proactive interactions in German.

5.3 Discussion of Demographic and Qualitative Influence Factors

Preferences for prompt formulations could not be found for all parameters and demographic groups. Nonetheless, as can be seen in Table 5 and Appendix C, conditional R^2 values indicate a high degree of subjectivity. These results point to interpersonal factors being a valid and intriguing research path to be considered further in future studies.

Differences were found for form of address and age, where younger participants rated "you" more positively than "I", while the effect was reversed for older participants who preferred referencing with "I". This points to differences regarding perceived agency, as well as needs around service-orientation. Younger study participants seem to appreciate an increased sense of agency, whereas older study participants preferred the service-orientation displayed by the VA's usage of

"I". This is in line with findings by Gollasch et al. [47] who found differing acceptance levels for proactive features between age groups. While younger study participants accepted proactivity, they still expressed a wish for interactions on demand. On the contrary, the age group of 35–65-year-olds preferred proactive suggestions over self-triggered conversations. These findings furthermore explain the age-related differences in preferences for long prompts. Age influenced subjects' preferences for sentence length in that older subjects rated long prompts as more intelligent than short prompts, while younger subjects displayed the exact opposite tendency. An increased acceptance for proactive suggestions in the age group of over 35-year-olds could go along with a higher acceptance for increased prompt lengths. Furthermore, a VA using longer prompts was probably perceived as being more knowledgeable, hence intelligent, by older study participants. For younger subjects with a higher wish for agency, intelligence was possibly better expressed through concise and condensed prompts.

One third of study participants did not perceive linguistic differences in the presented study prompts. Although lexical differences did not represent the majority of changes within prompts, most participants (~80%) attributed differences to lexis. This is in line with findings from Stier et al., who a) find a lack of linguistic awareness, and b) argue that this lack allows for intuitive prompt evaluations [132].

The general feedback we obtained supports previous findings around proactivity and the suitability of proactive use cases. Content, wording, suitability, and context were the most frequently mentioned influence factors for successful proactive interactions. Study participants demanded natural and informal wording, thereby supporting findings from Stier et al. [127]. Backing Schmidt et al.' findings [122–124], the content of proactive suggestions needs to entail all important information necessary to understand it, while keeping complexity low. As in Zargham et al.'s study [148], the suitability of proactivity depends on the usefulness and helpfulness of proactive suggestions, as well as on the applicability to the current driving scenario [127]. Playing into suitability considerations, study participants confirmed a fear of being overloaded and disturbed, which is discussed exhaustively in earlier works [5, 105, 114, 148]. These qualitative findings, while not new, underline the need for context-aware and context-sensitive proactive behavior with memory capacities around previous interactions. Furthermore, they stress the necessity for customizable and personalizable agent personas.

5.4 Limitations

While a plethora of syntactical, grammatical, as well as lexical parameters were considered in the present paper, it cannot raise a claim to completeness. Furthermore, proactive use cases outside the driving environment could come to different conclusions in terms of preferences for e.g., form of address. Demographic factors were found to have an influence on formulation preferences. Yet, clusters around demographic factors were not the focus of the study. As results point in a promising direction though, future research could concentrate more on demographics and potentially also include personality traits to gather further insights into decisive factors for formulation preferences. Furthermore, findings may be language-dependent and cannot necessarily be generalized from this study – which was conducted in German – to other languages. Still, the examined parameters can function as a starting point for future research in other languages. Lastly, the plain primary driving task and considerations around safety need discussing. Driving safety is paramount in in-car scenarios. However, as study participants did not experience safety-critical driving situations, concrete measures around safety were not part of the experimental setup. As suggested by related work, proactivity should not be triggered in case of demanding or safety-critical events [69, 105, 122, 127]. Still, the detection of these events may not always be reliable. As such, safety issues should be considered when designing proactive in-car features. We believe that the straightforward driving conditions enabled the detection of linguistic prompt preferences. As a low degree of syntactic complexity was preferred within simple driving conditions, it must not be assumed that these preferences will change in more demanding or safety-critical scenarios. We argue that testing prompts under complex driving conditions would have disregarded the large body of research stating that proactivity is unsuitable in the presence of demanding primary tasks. Nonetheless, we acknowledge that changes in primary task demand and security-relevance potentially influence linguistic preferences.

6 FUTURE WORK & CONCLUSION

While the present study showed linguistic preferences for the formulation of proactive prompts, it needs mentioning that these preferences were not found for all Likert scale dimensions. Future research should continue working on validating questionnaires especially tailored to the evaluation of single prompts, as current ones, like e.g., the UEQ, UEQ+, SASSI are assessing speech systems as a whole [57, 67, 76]. Studies show the importance of correct timing and relevance for a well-rounded user experience for proactive use cases. As the focus of this study lay in the formulation of prompts, the timing of proactivity was of secondary importance. Combining both how and when to proactively interrupt users poses an interesting research area for future work. Furthermore, varying the degree of driving complexity or security-relevance may allow the derivation of additional and more nuanced linguistic design guidelines.

While interruptibility as well as opportune moments have been researched broadly, thereby covering *when* to proactively interrupt users, research on *how* to interrupt them has received less attention. The present paper offers evidence regarding the importance and the practicability of this *how*, by providing concrete linguistic design guidelines for designing proactive prompts under consideration of demographic factors. A driving simulator study was carried out to determine formulation preferences in proactive situations in the vehicle. Indeed, preferences for proactive prompts were found on syntactical, grammatical, and lexical levels. A low level of complexity as well as a suggestive rather than an imposing speaking style were thereby found to be preferred by study participants. Still, formulation preferences for proactive prompts are not as pronounced as formulation preferences in other types of conversations with VAs. These findings underline that the existing design framework for proactive interactions needs to be enhanced to consider linguistic regards. However, they also show that the acceptance of proactive features is largely dependent on when users are interrupted and what they are interrupted with.

APPENDIX A: OVERVIEW OVER STUDY PROMPTS[18]

SYNTAX: SENTENCE STRUCTURE

parataxes | LIX: very low (<40)	*hypotaxes* | LIX: very low (<40)	*MCS* | LIX: low (<50)
Just for your information! There is much more traffic on your usual route today. It will take you about 10 minutes longer than usual. However, I have found a faster route for you. Would you like to take it?	Just for your information! There is much more traffic on your usual route today and it will take you about 10 minutes longer than usual. However, I have found a faster route for you. Would you like to take it?	Just for your information! There is a lot more traffic on your usual route today and it will take you about 10 minutes longer than usual, however, I have found a faster route for you. Would you like to take it?

SYNTAX: SENTENCE LENGTH

short | LIX: low (<50)	*medium* | LIX: low (<50)	*long* | LIX: low (<50)
Just for your information! The parking situation at the destination is moderate. Should I open the parking menu to show you suitable parking options?	Just for your information! I just saw that the parking situation at the destination looks rather moderate. Should I open the parking menu to show you suitable parking options?	Just for your information! We're almost there! I just saw that the parking situation at your destination looks rather moderate. Should I open the parking menu to show you suitable parking options?

SYNTAX: POSITION OF SUB-CLAUSES

prepositive | LIX: medium (<55)	*postpositive* | LIX: medium (<55)
Hey! If there are delays on your usual route, I'll look for a faster route for you. Would you like me to look for a faster route?	Hey! I'll look for a faster route for you if there are delays on your usual route. Would you like me to look for a faster route?

LEXIS: FORM OF ADDRESS

1st person sg "I" | LIX: medium (<55)	*2nd person sg "you"* | LIX: medium (<55)	*1st person pl "we"* | LIX: medium (<55)
Hey! Your map view can be customized to your personal preferences. Should I activate the customization?	Hey! Your map view can be customized to your personal preferences. Do you want to activate the customization?	Hey! Your map view can be customized to your personal preferences. Should we activate the customization?

LEXIS: POLITENESS

no politeness | LIX: medium (<55)	*politeness* | LIX: medium (<55)
Hey! I can switch on a relaxing atmosphere for you. Shall I?	Hey! I can gladly switch on a relaxing atmosphere for you. Shall I?

GRAMMAR: VOICE

active voice | LIX: low (<50)	*passive voice* | LIX: low (<50)
Hey! We will not make it to our destination with the remaining fuel range. Should I calculate a charging-optimized route for you?	Hey! We will not make it to our destination with the remaining fuel range. Should a charging-optimized route be calculated for you?

[18] The study was conducted in German. Prompts were translated by the author.

APPENDIX B: MEANS AND STANDARD DEVIATIONS ACROSS PARAMETERS

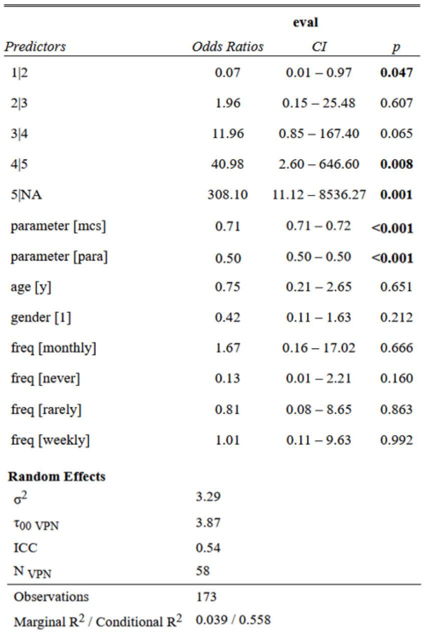
parameter		natural	positive	comprehensible	intelligent
hypotaxes	mean	2.45	2.10	2.22	5.16
	sd	1.06	1.17	1.11	1.06
parataxes	mean	2.30	1.91	2.37	1.93
	sd	1.16	0.99	1.29	0.84
MCS	mean	2.47	2.05	2.28	5.02
	sd	1.23	1.08	1.10	0.91
I	mean	2.81	2.64	3.38	5.55
	sd	1.00	1.07	1.55	1.06
you	mean	3.11	5.86	3.63	1.47
	sd	1.40	1.29	1.70	1.09
we	mean	2.72	5.30	2.65	5.03
	sd	3.24	2.78	3.41	2.67
short	mean	2.44	2.12	2.04	2.14
	sd	1.18	1.13	1.15	1.17
medium	mean	2.51	2.11	2.28	5.35
	sd	1.39	1.06	1.19	1.22
long	mean	2.66	2.21	2.21	5.14
	sd	1.34	1.09	1.18	1.10
prepositive	mean	2.69	2.29	2.48	2.14
	sd	1.35	1.08	1.25	0.78
postpositive	mean	2.65	2.33	2.68	5.30
	sd	1.19	0.89	1.28	0.91
politeness	mean	5.90	2.86	2.58	2.19
	sd	1.53	1.20	1.15	1.06
no politeness	mean	3.20	2.72	2.75	5.98
	sd	1.20	1.10	1.09	1.09
active	mean	2.88	2.02	3.21	2.05
	sd	1.28	1.03	1.42	1.11
passive	mean	2.60	1.86	2.93	2.02
	sd	1.21	0.88	1.41	0.94

APPENDIX C: RESULT TABLES FOR CUMULATIVE LINK MIXED MODELS

Parameters marked in bold represent statistically significant results.

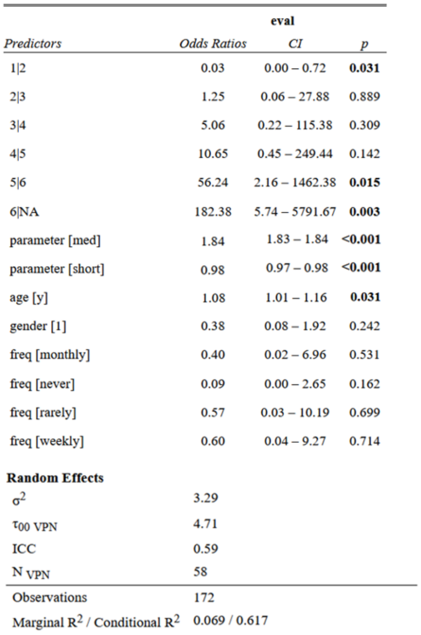
Intelligence

Sentence Structure



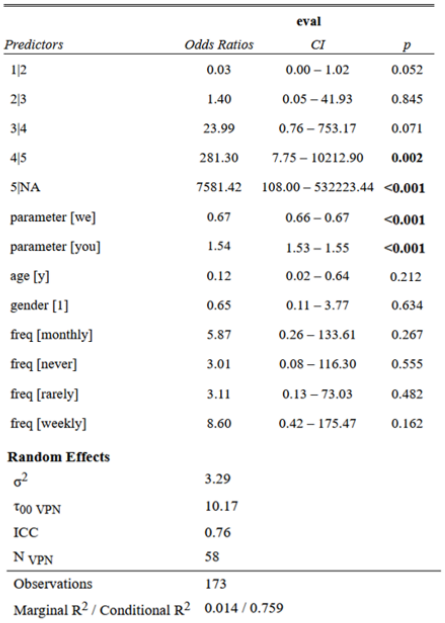
	eval		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.07	0.01 – 0.97	**0.047**
2|3	1.96	0.15 – 25.48	0.607
3|4	11.96	0.85 – 167.40	0.065
4|5	40.98	2.60 – 646.60	**0.008**
5|NA	308.10	11.12 – 8536.27	**0.001**
parameter [mcs]	0.71	0.71 – 0.72	**<0.001**
parameter [para]	0.50	0.50 – 0.50	**<0.001**
age [y]	0.75	0.21 – 2.65	0.651
gender [1]	0.42	0.11 – 1.63	0.212
freq [monthly]	1.67	0.16 – 17.02	0.666
freq [never]	0.13	0.01 – 2.21	0.160
freq [rarely]	0.81	0.08 – 8.65	0.863
freq [weekly]	1.01	0.11 – 9.63	0.992
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{VPN}}$	3.87		
ICC	0.54		
$\mathrm{N}_{\mathrm{VPN}}$	58		
Observations	173		
Marginal R^2 / Conditional R^2	0.039 / 0.558		

Sentence Length



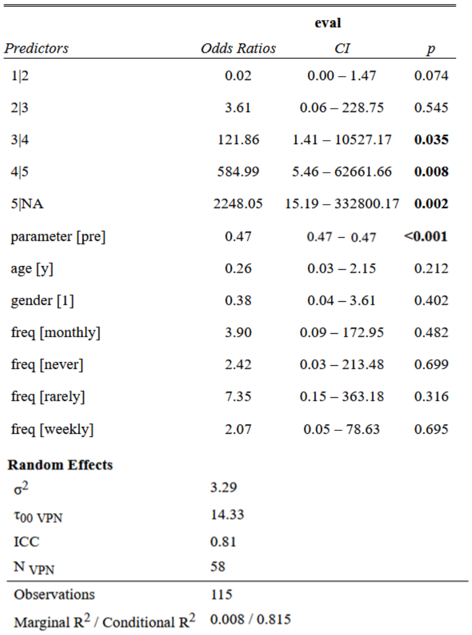
	eval		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.03	0.00 – 0.72	**0.031**
2|3	1.25	0.06 – 27.88	0.889
3|4	5.06	0.22 – 115.38	0.309
4|5	10.65	0.45 – 249.44	0.142
5|6	56.24	2.16 – 1462.38	**0.015**
6|NA	182.38	5.74 – 5791.67	**0.003**
parameter [med]	1.84	1.83 – 1.84	**<0.001**
parameter [short]	0.98	0.97 – 0.98	**<0.001**
age [y]	1.08	1.01 – 1.16	**0.031**
gender [1]	0.38	0.08 – 1.92	0.242
freq [monthly]	0.40	0.02 – 6.96	0.531
freq [never]	0.09	0.00 – 2.65	0.162
freq [rarely]	0.57	0.03 – 10.19	0.699
freq [weekly]	0.60	0.04 – 9.27	0.714
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{VPN}}$	4.71		
ICC	0.59		
$\mathrm{N}_{\mathrm{VPN}}$	58		
Observations	172		
Marginal R^2 / Conditional R^2	0.069 / 0.617		

Form of Address



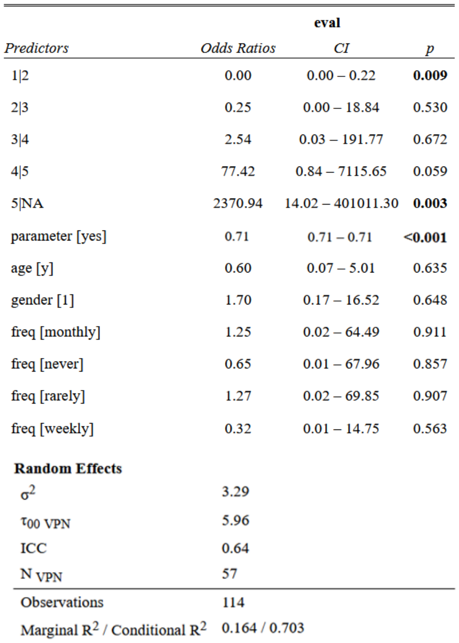
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.03	0.00 – 1.02	0.052
2|3	1.40	0.05 – 41.93	0.845
3|4	23.99	0.76 – 753.17	0.071
4|5	281.30	7.75 – 10212.90	**0.002**
5|NA	7581.42	108.00 – 532223.44	**<0.001**
parameter [we]	0.67	0.66 – 0.67	**<0.001**
parameter [you]	1.54	1.53 – 1.55	**<0.001**
age [y]	0.12	0.02 – 0.64	0.212
gender [1]	0.65	0.11 – 3.77	0.634
freq [monthly]	5.87	0.26 – 133.61	0.267
freq [never]	3.01	0.08 – 116.30	0.555
freq [rarely]	3.11	0.13 – 73.03	0.482
freq [weekly]	8.60	0.42 – 175.47	0.162
Random Effects			
σ^2	3.29		
τ_{00} VPN	10.17		
ICC	0.76		
N VPN	58		
Observations	173		
Marginal R^2 / Conditional R^2	0.014 / 0.759		

Position of Sub-Clauses



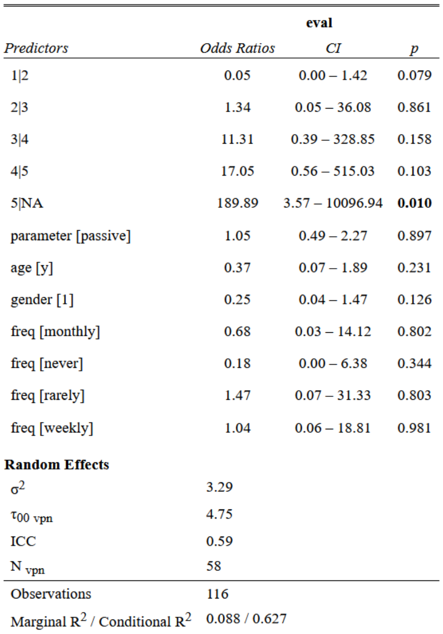
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.02	0.00 – 1.47	0.074
2|3	3.61	0.06 – 228.75	0.545
3|4	121.86	1.41 – 10527.17	**0.035**
4|5	584.99	5.46 – 62661.66	**0.008**
5|NA	2248.05	15.19 – 332800.17	**0.002**
parameter [pre]	0.47	0.47 – 0.47	**<0.001**
age [y]	0.26	0.03 – 2.15	0.212
gender [1]	0.38	0.04 – 3.61	0.402
freq [monthly]	3.90	0.09 – 172.95	0.482
freq [never]	2.42	0.03 – 213.48	0.699
freq [rarely]	7.35	0.15 – 363.18	0.316
freq [weekly]	2.07	0.05 – 78.63	0.695
Random Effects			
σ^2	3.29		
τ_{00} VPN	14.33		
ICC	0.81		
N VPN	58		
Observations	115		
Marginal R^2 / Conditional R^2	0.008 / 0.815		

Politeness



		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.00	0.00 – 0.22	**0.009**
2|3	0.25	0.00 – 18.84	0.530
3|4	2.54	0.03 – 191.77	0.672
4|5	77.42	0.84 – 7115.65	0.059
5|NA	2370.94	14.02 – 401011.30	**0.003**
parameter [yes]	0.71	0.71 – 0.71	**<0.001**
age [y]	0.60	0.07 – 5.01	0.635
gender [1]	1.70	0.17 – 16.52	0.648
freq [monthly]	1.25	0.02 – 64.49	0.911
freq [never]	0.65	0.01 – 67.96	0.857
freq [rarely]	1.27	0.02 – 69.85	0.907
freq [weekly]	0.32	0.01 – 14.75	0.563
Random Effects			
σ^2	3.29		
τ_{00} VPN	5.96		
ICC	0.64		
N VPN	57		
Observations	114		
Marginal R^2 / Conditional R^2	0.164 / 0.703		

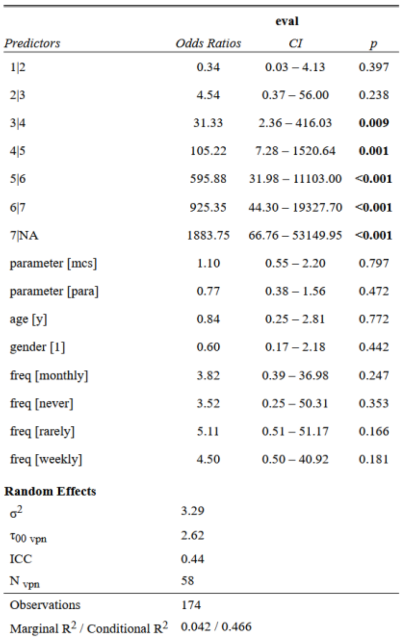
Mood



		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.05	0.00 – 1.42	0.079
2|3	1.34	0.05 – 36.08	0.861
3|4	11.31	0.39 – 328.85	0.158
4|5	17.05	0.56 – 515.03	0.103
5|NA	189.89	3.57 – 10096.94	**0.010**
parameter [passive]	1.05	0.49 – 2.27	0.897
age [y]	0.37	0.07 – 1.89	0.231
gender [1]	0.25	0.04 – 1.47	0.126
freq [monthly]	0.68	0.03 – 14.12	0.802
freq [never]	0.18	0.00 – 6.38	0.344
freq [rarely]	1.47	0.07 – 31.33	0.803
freq [weekly]	1.04	0.06 – 18.81	0.981
Random Effects			
σ^2	3.29		
τ_{00} vpn	4.75		
ICC	0.59		
N vpn	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.088 / 0.627		

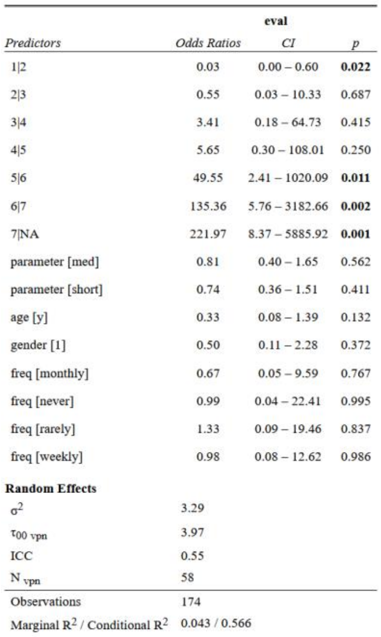
Naturalness

Sentence Structure



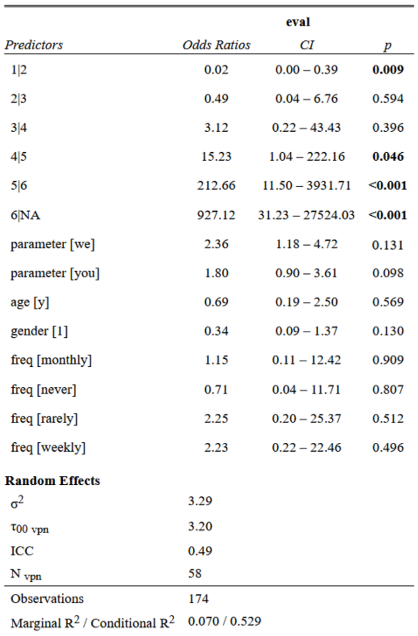
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.34	0.03 – 4.13	0.397
2|3	4.54	0.37 – 56.00	0.238
3|4	31.33	2.36 – 416.03	**0.009**
4|5	105.22	7.28 – 1520.64	**0.001**
5|6	595.88	31.98 – 11103.00	**<0.001**
6|7	925.35	44.30 – 19327.70	**<0.001**
7|NA	1883.75	66.76 – 53149.95	**<0.001**
parameter [mcs]	1.10	0.55 – 2.20	0.797
parameter [para]	0.77	0.38 – 1.56	0.472
age [y]	0.84	0.25 – 2.81	0.772
gender [1]	0.60	0.17 – 2.18	0.442
freq [monthly]	3.82	0.39 – 36.98	0.247
freq [never]	3.52	0.25 – 50.31	0.353
freq [rarely]	5.11	0.51 – 51.17	0.166
freq [weekly]	4.50	0.50 – 40.92	0.181
Random Effects			
σ^2	3.29		
$\tau_{00\ vpn}$	2.62		
ICC	0.44		
N $_{vpn}$	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.042 / 0.466		

Sentence Length



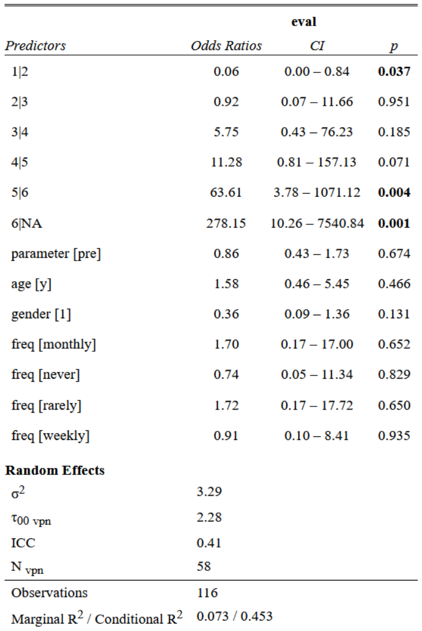
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.03	0.00 – 0.60	**0.022**
2|3	0.55	0.03 – 10.33	0.687
3|4	3.41	0.18 – 64.73	0.415
4|5	5.65	0.30 – 108.01	0.250
5|6	49.55	2.41 – 1020.09	**0.011**
6|7	135.36	5.76 – 3182.66	**0.002**
7|NA	221.97	8.37 – 5885.92	**0.001**
parameter [med]	0.81	0.40 – 1.65	0.562
parameter [short]	0.74	0.36 – 1.51	0.411
age [y]	0.33	0.08 – 1.39	0.132
gender [1]	0.50	0.11 – 2.28	0.372
freq [monthly]	0.67	0.05 – 9.59	0.767
freq [never]	0.99	0.04 – 22.41	0.995
freq [rarely]	1.33	0.09 – 19.46	0.837
freq [weekly]	0.98	0.08 – 12.62	0.986
Random Effects			
σ^2	3.29		
$\tau_{00\ vpn}$	3.97		
ICC	0.55		
N $_{vpn}$	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.043 / 0.566		

Form of Address



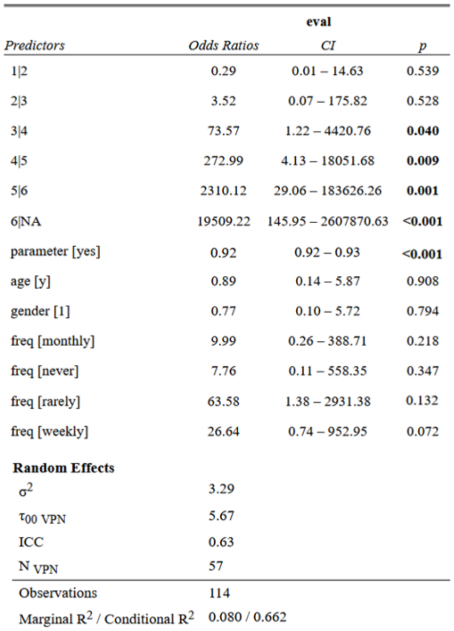
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.02	0.00 – 0.39	**0.009**
2|3	0.49	0.04 – 6.76	0.594
3|4	3.12	0.22 – 43.43	0.396
4|5	15.23	1.04 – 222.16	**0.046**
5|6	212.66	11.50 – 3931.71	**<0.001**
6|NA	927.12	31.23 – 27524.03	**<0.001**
parameter [we]	2.36	1.18 – 4.72	0.131
parameter [you]	1.80	0.90 – 3.61	0.098
age [y]	0.69	0.19 – 2.50	0.569
gender [1]	0.34	0.09 – 1.37	0.130
freq [monthly]	1.15	0.11 – 12.42	0.909
freq [never]	0.71	0.04 – 11.71	0.807
freq [rarely]	2.25	0.20 – 25.37	0.512
freq [weekly]	2.23	0.22 – 22.46	0.496
Random Effects			
σ^2	3.29		
$\tau_{00\ vpn}$	3.20		
ICC	0.49		
N $_{vpn}$	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.070 / 0.529		

Position of Sub-Clauses



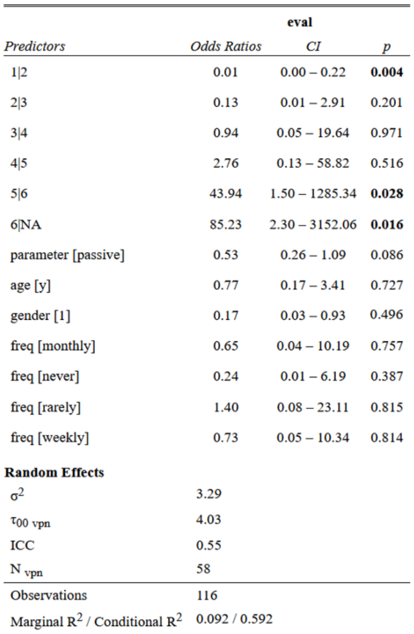
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.06	0.00 – 0.84	**0.037**
2|3	0.92	0.07 – 11.66	0.951
3|4	5.75	0.43 – 76.23	0.185
4|5	11.28	0.81 – 157.13	0.071
5|6	63.61	3.78 – 1071.12	**0.004**
6|NA	278.15	10.26 – 7540.84	**0.001**
parameter [pre]	0.86	0.43 – 1.73	0.674
age [y]	1.58	0.46 – 5.45	0.466
gender [1]	0.36	0.09 – 1.36	0.131
freq [monthly]	1.70	0.17 – 17.00	0.652
freq [never]	0.74	0.05 – 11.34	0.829
freq [rarely]	1.72	0.17 – 17.72	0.650
freq [weekly]	0.91	0.10 – 8.41	0.935
Random Effects			
σ^2	3.29		
$\tau_{00\ vpn}$	2.28		
ICC	0.41		
N $_{vpn}$	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.073 / 0.453		

Politeness



Predictors	Odds Ratios	CI	p
	eval		
1|2	0.29	0.01 – 14.63	0.539
2|3	3.52	0.07 – 175.82	0.528
3|4	73.57	1.22 – 4420.76	**0.040**
4|5	272.99	4.13 – 18051.68	**0.009**
5|6	2310.12	29.06 – 183626.26	**0.001**
6|NA	19509.22	145.95 – 2607870.63	**<0.001**
parameter [yes]	0.92	0.92 – 0.93	**<0.001**
age [y]	0.89	0.14 – 5.87	0.908
gender [1]	0.77	0.10 – 5.72	0.794
freq [monthly]	9.99	0.26 – 388.71	0.218
freq [never]	7.76	0.11 – 558.35	0.347
freq [rarely]	63.58	1.38 – 2931.38	0.132
freq [weekly]	26.64	0.74 – 952.95	0.072
Random Effects			
σ^2	3.29		
τ_{00} VPN	5.67		
ICC	0.63		
N VPN	57		
Observations	114		
Marginal R^2 / Conditional R^2	0.080 / 0.662		

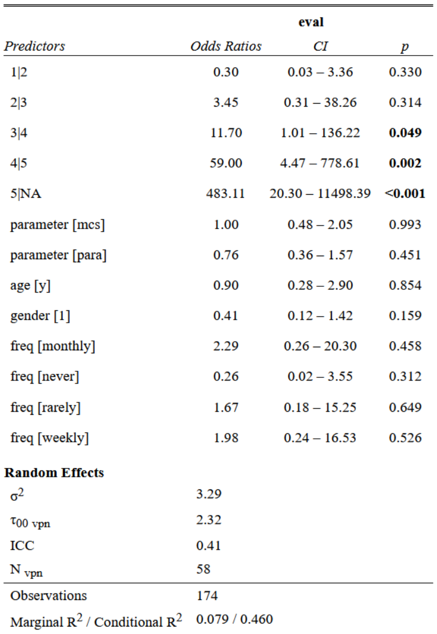
Mood



Predictors	Odds Ratios	CI	p
	eval		
1|2	0.01	0.00 – 0.22	**0.004**
2|3	0.13	0.01 – 2.91	0.201
3|4	0.94	0.05 – 19.64	0.971
4|5	2.76	0.13 – 58.82	0.516
5|6	43.94	1.50 – 1285.34	**0.028**
6|NA	85.23	2.30 – 3152.06	**0.016**
parameter [passive]	0.53	0.26 – 1.09	0.086
age [y]	0.77	0.17 – 3.41	0.727
gender [1]	0.17	0.03 – 0.93	0.496
freq [monthly]	0.65	0.04 – 10.19	0.757
freq [never]	0.24	0.01 – 6.19	0.387
freq [rarely]	1.40	0.08 – 23.11	0.815
freq [weekly]	0.73	0.05 – 10.34	0.814
Random Effects			
σ^2	3.29		
τ_{00} vpn	4.03		
ICC	0.55		
N vpn	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.092 / 0.592		

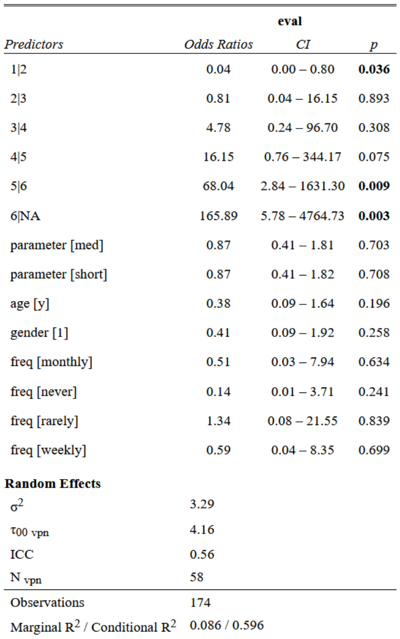
Positivity

Sentence Structure



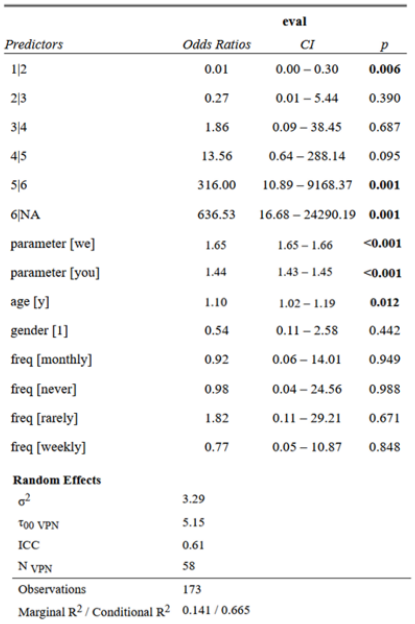
Predictors	Odds Ratios	CI	p
	eval		
1|2	0.30	0.03 – 3.36	0.330
2|3	3.45	0.31 – 38.26	0.314
3|4	11.70	1.01 – 136.22	**0.049**
4|5	59.00	4.47 – 778.61	**0.002**
5|NA	483.11	20.30 – 11498.39	**<0.001**
parameter [mcs]	1.00	0.48 – 2.05	0.993
parameter [para]	0.76	0.36 – 1.57	0.451
age [y]	0.90	0.28 – 2.90	0.854
gender [1]	0.41	0.12 – 1.42	0.159
freq [monthly]	2.29	0.26 – 20.30	0.458
freq [never]	0.26	0.02 – 3.55	0.312
freq [rarely]	1.67	0.18 – 15.25	0.649
freq [weekly]	1.98	0.24 – 16.53	0.526
Random Effects			
σ^2	3.29		
τ_{00} vpn	2.32		
ICC	0.41		
N vpn	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.079 / 0.460		

Sentence Length



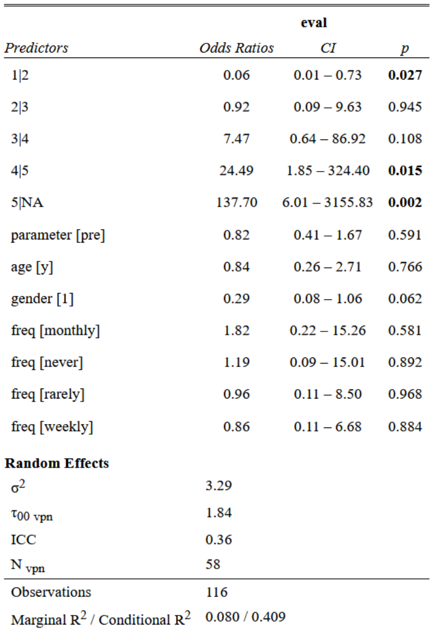
Predictors	Odds Ratios	CI	p
	eval		
1|2	0.04	0.00 – 0.80	**0.036**
2|3	0.81	0.04 – 16.15	0.893
3|4	4.78	0.24 – 96.70	0.308
4|5	16.15	0.76 – 344.17	0.075
5|6	68.04	2.84 – 1631.30	**0.009**
6|NA	165.89	5.78 – 4764.73	**0.003**
parameter [med]	0.87	0.41 – 1.81	0.703
parameter [short]	0.87	0.41 – 1.82	0.708
age [y]	0.38	0.09 – 1.64	0.196
gender [1]	0.41	0.09 – 1.92	0.258
freq [monthly]	0.51	0.03 – 7.94	0.634
freq [never]	0.14	0.01 – 3.71	0.241
freq [rarely]	1.34	0.08 – 21.55	0.839
freq [weekly]	0.59	0.04 – 8.35	0.699
Random Effects			
σ^2	3.29		
τ_{00} vpn	4.16		
ICC	0.56		
N vpn	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.086 / 0.596		

Form of Address



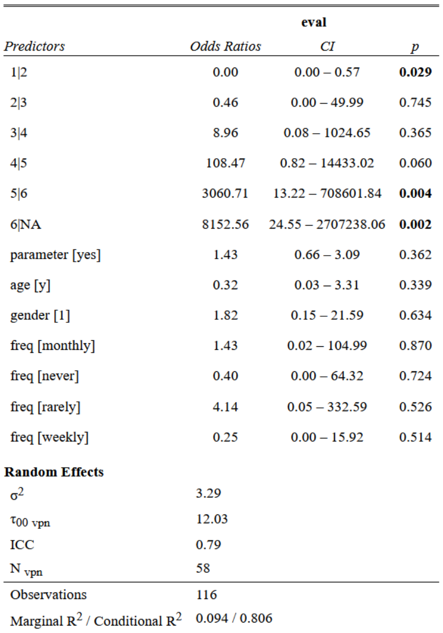
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.01	0.00 – 0.30	**0.006**
2|3	0.27	0.01 – 5.44	0.390
3|4	1.86	0.09 – 38.45	0.687
4|5	13.56	0.64 – 288.14	0.095
5|6	316.00	10.89 – 9168.37	**0.001**
6|NA	636.53	16.68 – 24290.19	**0.001**
parameter [we]	1.65	1.65 – 1.66	**<0.001**
parameter [you]	1.44	1.43 – 1.45	**<0.001**
age [y]	1.10	1.02 – 1.19	**0.012**
gender [1]	0.54	0.11 – 2.58	0.442
freq [monthly]	0.92	0.06 – 14.01	0.949
freq [never]	0.98	0.04 – 24.56	0.988
freq [rarely]	1.82	0.11 – 29.21	0.671
freq [weekly]	0.77	0.05 – 10.87	0.848
Random Effects			
σ^2	3.29		
τ_{00} VPN	5.15		
ICC	0.61		
N VPN	58		
Observations	173		
Marginal R^2 / Conditional R^2	0.141 / 0.665		

Position of Sub-Clauses



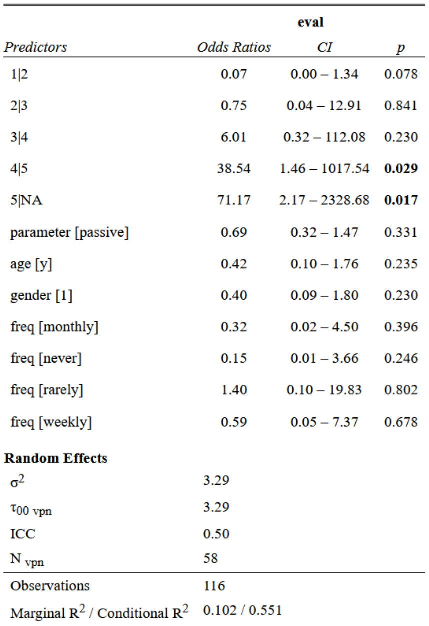
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.06	0.01 – 0.73	**0.027**
2|3	0.92	0.09 – 9.63	0.945
3|4	7.47	0.64 – 86.92	0.108
4|5	24.49	1.85 – 324.40	**0.015**
5|NA	137.70	6.01 – 3155.83	**0.002**
parameter [pre]	0.82	0.41 – 1.67	0.591
age [y]	0.84	0.26 – 2.71	0.766
gender [1]	0.29	0.08 – 1.06	0.062
freq [monthly]	1.82	0.22 – 15.26	0.581
freq [never]	1.19	0.09 – 15.01	0.892
freq [rarely]	0.96	0.11 – 8.50	0.968
freq [weekly]	0.86	0.11 – 6.68	0.884
Random Effects			
σ^2	3.29		
τ_{00} vpn	1.84		
ICC	0.36		
N vpn	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.080 / 0.409		

Politeness



		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.00	0.00 – 0.57	**0.029**
2|3	0.46	0.00 – 49.99	0.745
3|4	8.96	0.08 – 1024.65	0.365
4|5	108.47	0.82 – 14433.02	0.060
5|6	3060.71	13.22 – 708601.84	**0.004**
6|NA	8152.56	24.55 – 2707238.06	**0.002**
parameter [yes]	1.43	0.66 – 3.09	0.362
age [y]	0.32	0.03 – 3.31	0.339
gender [1]	1.82	0.15 – 21.59	0.634
freq [monthly]	1.43	0.02 – 104.99	0.870
freq [never]	0.40	0.00 – 64.32	0.724
freq [rarely]	4.14	0.05 – 332.59	0.526
freq [weekly]	0.25	0.00 – 15.92	0.514
Random Effects			
σ^2	3.29		
τ_{00} vpn	12.03		
ICC	0.79		
N vpn	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.094 / 0.806		

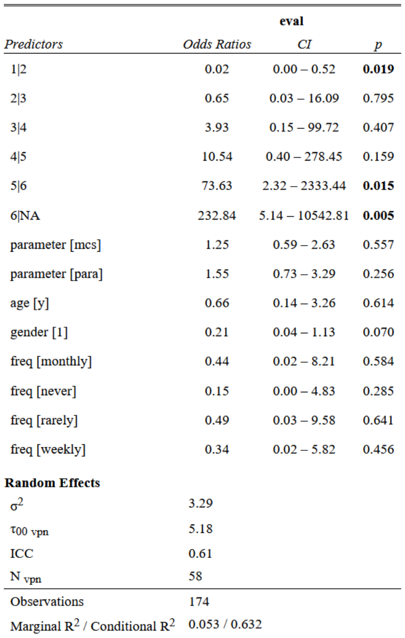
Mood



		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.07	0.00 – 1.34	0.078
2|3	0.75	0.04 – 12.91	0.841
3|4	6.01	0.32 – 112.08	0.230
4|5	38.54	1.46 – 1017.54	**0.029**
5|NA	71.17	2.17 – 2328.68	**0.017**
parameter [passive]	0.69	0.32 – 1.47	0.331
age [y]	0.42	0.10 – 1.76	0.235
gender [1]	0.40	0.09 – 1.80	0.230
freq [monthly]	0.32	0.02 – 4.50	0.396
freq [never]	0.15	0.01 – 3.66	0.246
freq [rarely]	1.40	0.10 – 19.83	0.802
freq [weekly]	0.59	0.05 – 7.37	0.678
Random Effects			
σ^2	3.29		
τ_{00} vpn	3.29		
ICC	0.50		
N vpn	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.102 / 0.551		

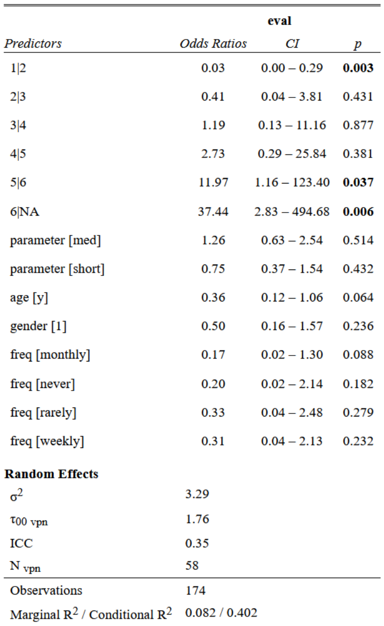
Comprehensibility

Sentence Structure



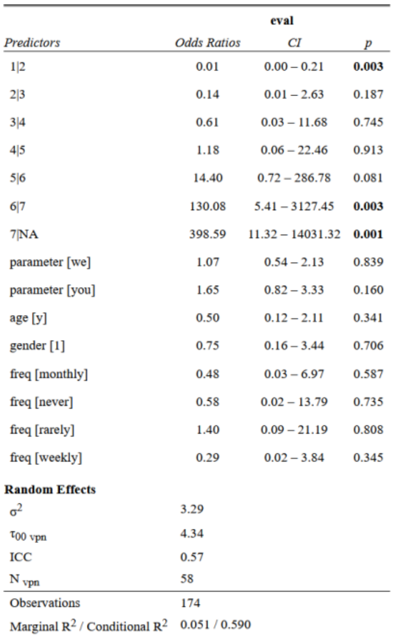
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.02	0.00 – 0.52	**0.019**
2|3	0.65	0.03 – 16.09	0.795
3|4	3.93	0.15 – 99.72	0.407
4|5	10.54	0.40 – 278.45	0.159
5|6	73.63	2.32 – 2333.44	**0.015**
6|NA	232.84	5.14 – 10542.81	**0.005**
parameter [mcs]	1.25	0.59 – 2.63	0.557
parameter [para]	1.55	0.73 – 3.29	0.256
age [y]	0.66	0.14 – 3.26	0.614
gender [1]	0.21	0.04 – 1.13	0.070
freq [monthly]	0.44	0.02 – 8.21	0.584
freq [never]	0.15	0.00 – 4.83	0.285
freq [rarely]	0.49	0.03 – 9.58	0.641
freq [weekly]	0.34	0.02 – 5.82	0.456
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{vpn}}$	5.18		
ICC	0.61		
N_{vpn}	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.053 / 0.632		

Sentence Length



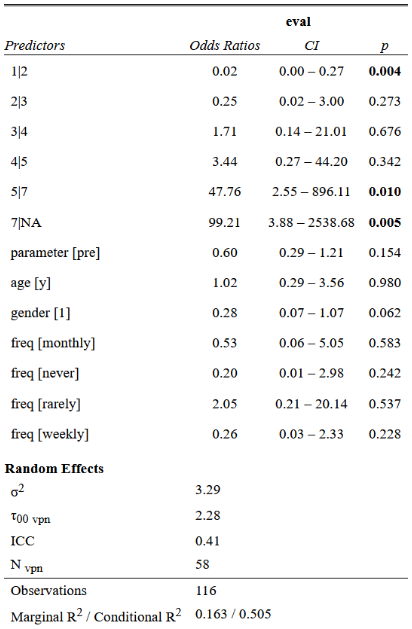
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.03	0.00 – 0.29	**0.003**
2|3	0.41	0.04 – 3.81	0.431
3|4	1.19	0.13 – 11.16	0.877
4|5	2.73	0.29 – 25.84	0.381
5|6	11.97	1.16 – 123.40	**0.037**
6|NA	37.44	2.83 – 494.68	**0.006**
parameter [med]	1.26	0.63 – 2.54	0.514
parameter [short]	0.75	0.37 – 1.54	0.432
age [y]	0.36	0.12 – 1.06	0.064
gender [1]	0.50	0.16 – 1.57	0.236
freq [monthly]	0.17	0.02 – 1.30	0.088
freq [never]	0.20	0.02 – 2.14	0.182
freq [rarely]	0.33	0.04 – 2.48	0.279
freq [weekly]	0.31	0.04 – 2.13	0.232
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{vpn}}$	1.76		
ICC	0.35		
N_{vpn}	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.082 / 0.402		

Form of Address



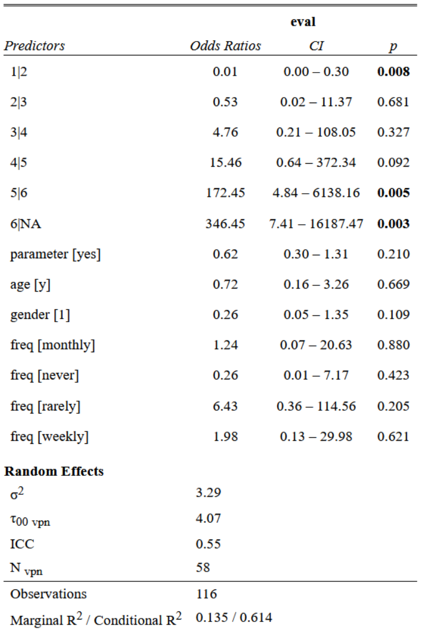
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.01	0.00 – 0.21	**0.003**
2|3	0.14	0.01 – 2.63	0.187
3|4	0.61	0.03 – 11.68	0.745
4|5	1.18	0.06 – 22.46	0.913
5|6	14.40	0.72 – 286.78	0.081
6|7	130.08	5.41 – 3127.45	**0.003**
7|NA	398.59	11.32 – 14031.32	**0.001**
parameter [we]	1.07	0.54 – 2.13	0.839
parameter [you]	1.65	0.82 – 3.33	0.160
age [y]	0.50	0.12 – 2.11	0.341
gender [1]	0.75	0.16 – 3.44	0.706
freq [monthly]	0.48	0.03 – 6.97	0.587
freq [never]	0.58	0.02 – 13.79	0.735
freq [rarely]	1.40	0.09 – 21.19	0.808
freq [weekly]	0.29	0.02 – 3.84	0.345
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{vpn}}$	4.34		
ICC	0.57		
N_{vpn}	58		
Observations	174		
Marginal R^2 / Conditional R^2	0.051 / 0.590		

Position of Sub-Clauses



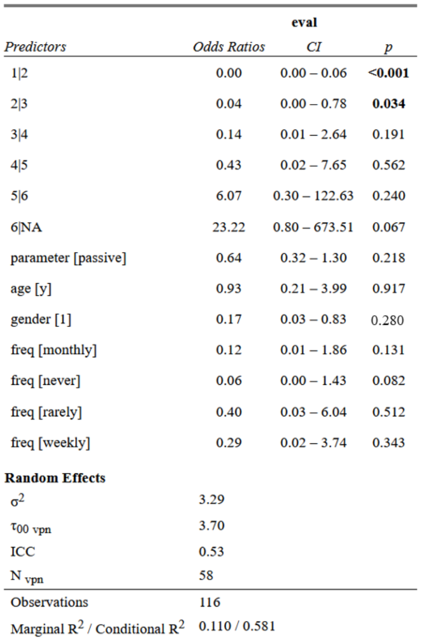
		eval	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.02	0.00 – 0.27	**0.004**
2|3	0.25	0.02 – 3.00	0.273
3|4	1.71	0.14 – 21.01	0.676
4|5	3.44	0.27 – 44.20	0.342
5|7	47.76	2.55 – 896.11	**0.010**
7|NA	99.21	3.88 – 2538.68	**0.005**
parameter [pre]	0.60	0.29 – 1.21	0.154
age [y]	1.02	0.29 – 3.56	0.980
gender [1]	0.28	0.07 – 1.07	0.062
freq [monthly]	0.53	0.06 – 5.05	0.583
freq [never]	0.20	0.01 – 2.98	0.242
freq [rarely]	2.05	0.21 – 20.14	0.537
freq [weekly]	0.26	0.03 – 2.33	0.228
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{vpn}}$	2.28		
ICC	0.41		
N_{vpn}	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.163 / 0.505		

Politeness



	eval		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.01	0.00 – 0.30	**0.008**
2|3	0.53	0.02 – 11.37	0.681
3|4	4.76	0.21 – 108.05	0.327
4|5	15.46	0.64 – 372.34	0.092
5|6	172.45	4.84 – 6138.16	**0.005**
6|NA	346.45	7.41 – 16187.47	**0.003**
parameter [yes]	0.62	0.30 – 1.31	0.210
age [y]	0.72	0.16 – 3.26	0.669
gender [1]	0.26	0.05 – 1.35	0.109
freq [monthly]	1.24	0.07 – 20.63	0.880
freq [never]	0.26	0.01 – 7.17	0.423
freq [rarely]	6.43	0.36 – 114.56	0.205
freq [weekly]	1.98	0.13 – 29.98	0.621
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{vpn}}$	4.07		
ICC	0.55		
N_{vpn}	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.135 / 0.614		

Mood



	eval		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.00	0.00 – 0.06	**<0.001**
2|3	0.04	0.00 – 0.78	**0.034**
3|4	0.14	0.01 – 2.64	0.191
4|5	0.43	0.02 – 7.65	0.562
5|6	6.07	0.30 – 122.63	0.240
6|NA	23.22	0.80 – 673.51	0.067
parameter [passive]	0.64	0.32 – 1.30	0.218
age [y]	0.93	0.21 – 3.99	0.917
gender [1]	0.17	0.03 – 0.83	0.280
freq [monthly]	0.12	0.01 – 1.86	0.131
freq [never]	0.06	0.00 – 1.43	0.082
freq [rarely]	0.40	0.03 – 6.04	0.512
freq [weekly]	0.29	0.02 – 3.74	0.343
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{vpn}}$	3.70		
ICC	0.53		
N_{vpn}	58		
Observations	116		
Marginal R^2 / Conditional R^2	0.110 / 0.581		

V. FAILING WITH GRACE: EXPLORING THE ROLE OF REPAIR COSTS IN CONVERSATIONAL BREAKDOWNS WITH IN-CAR VOICE ASSISTANTS[19]

Technological advancements over the past years have led to a leap in Conversational User Interface (CUI) performance. However, interactions with CUIs are not error-free and error handling thus plays a crucial role in CUI design. So far, CUI repair strategies oftentimes neglect what is known from Human-Human Interaction (HHI). The present work borrows from human-centered error handling and applies the principle of Least Collaborative Effort to Human-Computer Interaction (HCI). A within-subjects WoZ driving simulator study (n=48) is conducted to examine user preferences qualitatively and quantitatively for seven error handling strategies in an automotive setting. Our findings show that human-centered repair strategies comply with the principle of Least Collaborative Effort and lead to significantly lower repair costs than a classic HCI error handling strategy. These results introduce costs as a measurement tool for examining strategy preferences when repairing erroneous dialogs and offer assistance for designing user-centric error handling strategies.

1 INTRODUCTION

Communication is a most complex undertaking and requires the precise and detailed accordance of communication partners, knowledge, timing, and resources. To lead successful dialogs, a collective and collaborative effort is needed from all parties involved [21, 130]. Even if these parties “are rational and cooperative, inhabit the same location, speak the same language, share much of the same knowledge, and use common wording, there is no guarantee that one will understand the other (…)” [21, p. 1]. Still, “there is never an unrecoverable error state” [53, p. 1] either when humans talk to each other. Human interlocutors have a variety of verbal and non-verbal cues and rich context at their disposal when repairing conversational errors. In HHI, recovering from errors can be facilitated not only through speech, but with the aid of gestures, facial expressions, gaze, and deixis [11]. Furthermore, conversational partners rely on so-called social signals to communicate their intentions [40]. Non-verbal social signals form a key component of

[19] As the leading author, I developed the research idea as well as the experiment design and conducted and analyzed all described studies. Dr. Christoph Draxler and Dr. Thurid Vogt supported the development of the research idea and the experiment design and provided feedback on the overall work.

social intelligence [139] and play a role in e.g., turn-taking behavior [46]. Combined, these elements make error handling a seamless and nearly subconscious act in HHI [16]. This does not hold true for interactions with CUIs: while humans repair errors multimodally, CUIs are restricted in that they typically lack non-verbal communication patterns and possibilities. Moreover, recovering from errors requires grounding, meaning the conformity of a given CUI's and a human interlocutor's mental model. Arriving at this conformity can be impeded by a legitimate lack of technical knowledge from the user's side, as well as a lack of figurative meaning and pragmatics from the CUI's side [15, 99].

The last years have shown a leap in CUI performance in terms of decreasing word error rates and improved intent recognition. Still, recognition and accuracy differ between CUIs and are oftentimes use case- and speaker-dependent [54]. Error handling hence still plays a crucial role in any CUI design strategy. If a dialog fails, the collaborative handling of errors is crucial to arrive at a common understanding. Oftentimes, CUIs do not provide transparent or concrete feedback on the type of an error though, which leaves users in the dark on how to successfully repair it [112]. Subsequently, users are found to engage in repair strategies such as hyper-articulation, simplification, talking more slowly, and providing additional information [9, 15, 99]. While these repair strategies work well in HHI, they are not directly applicable to HCI and rather drive users into disconcerting error spirals. As a consequence, user trust and satisfaction with CUIs can decrease and ultimately lead to product abandonment [79, 86, 130]. If users are conducting a primary task while using a CUI, effects of errors and the subsequent error handling strategy are even more far-reaching. A driving simulator study by Nass and Brave [101] found a negative impact of erroneous dialog on study participants' driving performance and their attention to the primary driving task. Voice is an especially intriguing modality for in-car contexts as it leads to fewer driving errors, less distraction and reduced cognitive load compared to touch [74]. Yet, designed improperly, CUI usage can even adversely affect traffic safety [6, 135]. Hence, a thoughtful and goal-oriented error handling is especially crucial in an in-car context.

When designing for HCI, looking at conversational equivalents in HHI is exceedingly helpful. Currently, HCI repair strategies are commonly restricted to generic error messages, e.g. *I didn't catch that*. Further strategies ask users to repeat commands, redirect them to tasks the system can handle, and elaborate on utterances that will be understood; in rare cases, designers fall back on humor [108, 111, 112]. HHI error handling strategies on the other hand explicitly label error sources, e.g., *I didn't catch that last word*, or *I did not understand that acoustically*. Furthermore, humans try to advance a dialog by continuing descriptions, asking task-related questions, and making assumptions regarding their conversational partner's intent – oftentimes without indicating errors [130].

HHI error handling strategies thereby follow the so-called principle of *Least Collaborative Effort*. Coined by Clark and Brennan [31], the principle extends Grice's [49] maxim of quantity and manner and expects conversational partners to minimize the effort or so-called *costs* going along with repairing a conversation. The researchers provide an extensive overview over conversational costs, reaching from straightforward formulation costs to psycholinguistic fault costs [31]. Formulation costs describe the number of words and conversation turns needed to repair an error. The more words and dialog turns are needed to correct an error, the higher formulation costs become. Conversational partners want to keep these costs as low as possible to adhere to Grice's [49] maxim of quantity and manner as well as to the principle of Least Collaborative Effort [31]. Fault costs on the other hand can make a given speaker "look foolish, illiterate, or impolite" [31, p. 145], and are avoided in order to prevent face-threatening acts [18]. Formulation costs present a most interesting starting point in determining the suitability of error handling strategies. To the best of our knowledge, costs as well as the principle of Least Collaborative Effort do not yet find resonance in HCI error handling strategies. However, we believe that borrowing from the principle of Least Collaborative Effort leads to a more user-centric, natural, and thus more successful error handling approach. Furthermore, cost calculations can conceivably prove useful as a measurement tool for surveying user preferences for error handling strategies. Although there are validated measurement instruments for assessing the user experience of entire systems [17, 57, 67], these measurement instruments are lacking and/or are not validated for evaluating individual dialogs and dialog steps. The present paper attempts to close this gap by compiling hands-on guidelines for designing human-centred error handling strategies. Furthermore, it aims at investigating whether cost calculations are an indicator for successful error handling strategies and whether the Principle of Least Collaborative Effort should thus be assigned a role in HCI.

The structure of this paper is as follows: chapter 2 reviews relevant literature regarding error handling in HHI and HCI. Based on these considerations, chapter 3 presents three HHI and cost informed error handling strategies for German task-oriented conversations in the vehicle. By means of a within-subjects driving simulator study, these error handling strategies are tested against a conventional HCI error handling strategy. In a subsequent qualitative post-interview, study participants are asked to rank all strategies regarding their appropriateness as in-car error handling strategies. In chapter 4, quantitative and qualitative analyses are conducted. Results are evaluated and discussed in chapter 5, where we show that costs are an easily applicable tool to predict preferences for error handling strategies. Based on these results, chapter 6 provides concrete design recommendations for CUI design and implementation. The paper closes with final conclusions in chapter 7. The following research questions and hypotheses will be answered throughout the paper:

RQ1: Do different error handling strategies vary regarding their repair costs?

RQ2: Do the costs associated with repairing errors influence the preference for error handling strategies?

Based on these research questions, two hypotheses are formulated:

H1: HHI informed error handling strategies lead to lower repair costs than a HCI error handling strategy.

H2: The error handling strategy associated with the lowest repair costs is preferred by study participants.

2 RELATED WORK

2.1 Error Handling in Human-Computer Interaction

CUIs are trained to expect certain input at a fixed iteration in a dialog. In an exemplary use case where a user would like to book a flight, a CUI requires information on departure and arrival as well as time, date, or personal information such as number of passengers. In case input is missing entirely (no input), not understood correctly (Automatic Speech Recognition (ASR) error) or not mappable to the current or a known intent (Natural Language Understanding (NLU) error), an error has occurred, and a suitable error handling strategy needs to be carried out. The selection of such a strategy thereby depends on the type of error. Non-understandings require different strategies than NLU or ASR errors. Depending on the level of recognition confidence, more or less conservative repair strategies are conceivable. A conservative approach could ask users to repeat their commands. A less conservative repair on the other hand could suggest solving an error or ignore it entirely, thereby running the risk of introducing a new error in the conversation.

Aneja et al. [3] summarize it well when stating that some “errors are likely to be considered more frustrating (…) than others, while some mistakes could potentially be viewed as cute, amusing, or make the system feel less threatening" [3, p. 1]. A study exploring the long-term behavior of Alexa users with a focus on communication breakdowns found that breakdowns only had negative effects on user satisfaction when they were not successfully repaired. Users hence tolerate errors as long as they can be solved [35, 89]. Still, errors need to be answered with dedicated design strategies as users are also found to abandon CUIs because their trust towards an erroneous system diminishes and they enjoy interactions less [4, 60, 86]. A study by Motta et al. [97] stresses the importance of supporting users during error handling by discovering that study participants willingly follow recovery paths prescribed by a CUI.

A question frequently studied in the realm of conversational failures is the effect of blame and apologies. Mahmood et al. [87] studied the impact of sincere apologies and the assignment of blame on user experience and agent perception. An agent's intelligence, likability, and the efficiency to recover from errors increased in case the agent accepted blame and apologized sincerely. This effect had a context-sensitive component though, as Ashktorab et al. [4] found that "acknowledging a mistake lowers the likability and perceived intelligence of the agent (…)" [4, p. 3] in task-oriented conversations. Blaming mistakes on the user on the other hand can have severe implications: Nass and Brave [101] found lower performance ratings, less system likability, and reduced attention to a primary task when users were blamed for errors. Moreover, study participants' driving performance was negatively impacted by erroneous dialogs in this study.

2.2 HCI Error Handling Strategies

The following paragraphs describe currently used HCI error handling strategies and their impact on user satisfaction and evaluation. Zargham et al. [147] developed a voice-controlled game and compared an anticipatory error handling strategy with the more traditional error handling strategy of asking users to repeat their utterances. In their study, they found that user experience as well as perceived intuitiveness of control in the game improved significantly when applying anticipatory error handling. Anticipatory error handling means not indicating an error and trying to advance a given dialog by presenting a reasonable next step. In case of non-matches between user intent and anticipatory action, usability declined in Zargham et al.'s study and left users feeling confused and deceived. Still, study participants in the anticipatory error handling condition reported lower numbers of errors than participants in the repetition-based baseline condition. The actual error rate did not differ significantly between conditions and was never smaller than 17% though. Due to the perceived lower number of errors, the anticipatory intervention group rated the voice game to be more intelligent than their baseline counterparts. The research team concludes that “error handling can significantly improve the usability of a speech-controlled video game” and that “false handling can impair both the experience as well as the learning progress” [147, p. 10–11]. Bohus and Rudnicky’s [12] research seconds that. In a study they conducted, the authors looked at ten non-understanding recovery strategies and compared their performance. Their results showed that advancing the conversation by ignoring errors and trying an alternative dialog plan performed best. Engelhardt et al. [41] add to this research by exploring different error handling strategies in conversational breakdowns in HRI. In their speech-based experiment, they compared three error handling strategies, namely ignoring errors, apologizing for errors, and solving errors collaboratively. The title of their research “Better Faulty than Sorry” gives away their findings: likeability and perceived intelligence decreased for the apology condition, while ignoring errors increased intelligence and animacy.

While research shows that errors in HCI lead to a decreased user experience, Aneja et al. [3] found that not all errors equally and foremost equally negatively impacted the perception of a conversational agent. The researchers analyzed five typical HCI conversational errors and their respective impact on an agent. Error handling strategies involving repetitions from the agent and subsequent clarifications from the user’s side were detrimental to the agent’s perceived intelligence. The agent’s overall likability furthermore decreased with an increase in turn-taking (i.e., the number of dialog turns in a conversation where one turn is one utterance plus one CUI prompt). Nonetheless, a positive impact on likability was recorded for coherence errors. The researchers interpret this finding as being “more in line with user expectations of a more approachable, human-like agent that will likely make logical mistakes (…)” [3, p. 6]. Anthropomorphism hence increased with coherence errors but decreased with repetitions. Mirnig et al. [96] obtained similar results in HRI, where interactions with a robot committing errors were rated significantly better than interactions with a robot which never failed. This research indicates that the interpretation of errors depends on the context and – while being detrimental to goal-oriented conversations – can increase likability of an embodied agent. Kontogiorgos et al. [71] emphasize this. In their study, users considered abandoning smart speakers in case of errors, while they would still interact with an erroneous embodied assistant.

Ashktorab et al. [4] explored different repair strategies for conversational breakdowns in chatbots, including asking users for confirmations, providing options, repeating user utterances, or giving no evidence of an error. The researchers differentiated between task-oriented and chit chat conversations and found suitable error handling to depend on the type of conversation. While CUIs should keep engaging users in case of chit chat use cases, a quick and swift error handling is necessary in task-oriented conversations. Providing options to correct an error turned out to be users' preferred error handling strategy. If a repair was possible within one dialog turn, users accepted error handling strategies which did not indicate errors directly. The researchers conclude that collaborative measures taken by an agent to handle errors should include the acknowledgement of errors as well as proactive suggestions to mitigate them [4].

2.3 User Handling of Conversational Breakdowns

While the previous paragraphs were concerned with the design of different error handling strategies, the following section addresses conversational breakdowns from a user perspective.

Repairing errors in CUIs can be a challenging and burdensome task for users. Especially in speech-only systems, the lack of visible aids in form of screens or embodied agents hinders discoverability and learnability [45, 99]. In HHI, conversational partners repair errors multimodally, meaning they enhance speech with non-verbal elements, gaze or gestures [31]. In HCI on the other hand, error handling can oftentimes only be facilitated via speech. A study by Suhm et al. [137] finds multimodal error handling to outperform unimodal error handling though. Initially, users prefer to trigger interactions via speech. As speech "is slow for presenting information, is transient and therefore difficult to review or edit" [129, p. 63], they favor making repairs via touch. While running on the pledge of natural language, recognition capacities and abilities frequently fall short of this promise [146]. CUIs are still oftentimes not tailored to the complex, inventive, and multi-facetted nature of human language [33, 34, 86, 112]. Due to their lack of understanding system boundaries and recognition principles, users cannot repair errors in a manner that facilitates understanding from the CUI's side [99]. Rather, they apply error handling strategies from HHI. Cheng et al. [27], who centred their research around error handling strategies adopted by children, support these findings. In their research, children utilized strategies from HHI to repair erroneous interactions with CUIs. These strategies were similar to error handling strategies applied by adults, although children were more persistent and patient when repairing interactions.

Beneteau et al. [8] add to these findings by stressing that CUIs need to provide collaborative error handling strategies between CUIs and their users, instead of leaving repairs to users alone.

Myers et al. [99] focused their study on strategies users apply to overcome obstacles. While most errors in CUIs stem from natural language processing (NLP), they found other errors to lead to elevated user frustration and confusion. When confronted with an error case, users across demographics engaged in hyper-articulation, simplifications, and oftentimes provided additional information. The strategy thereby partly depended on the CUI's perceived intelligence. The more intelligent the CUI was judged the more additional information users provided it with. These findings are mirrored by Motta et al. [97] who moreover found that the users' choice of repair strategy depended on the preceding CUI response. Errors in ASR were met with overpronounciation and repetitions, while wrongful task execution led to exploratory behavior. In Giuliani et al.'s [46] research, study participants' responses to erroneous interactions in HRI were similarly fine-grained. Error handling strategies depended on whether an error was categorized as a technical failure or a social norm violation. In case of technical errors, study participants repaired errors with fewer words than in case of social norm violations. By identifying technical limitations as source of the error, they adapted their speaking style to support the robot's perceived poorer language abilities [46]. Sensitivity to different error sources as well as tailored error handling strategies were also found by Kim et al. [64] in an in-car context. Study participants categorized and attributed errors to different sources, such as ASR errors, input errors, or system boundary errors and adapted their error handling strategies accordingly. If study participants suspected an ASR error, they repeated their initial utterance more slowly and condensed. Assuming a system boundary error, they reformulated their initial utterance and/or simplified it [64]. Mavrina et al. [89] additionally discovered that reformulations were predominantly used if users believed errors did not originate from their side.

While the repair strategies outlined above work well in HHI, they are not necessarily suitable for CUIs. Hyper-articulation means speaking more slowly, loudly, and clearly which does not reliably lead to better recognition. Even more so, as Myers et al. [99] found that users tend to hyperarticulate content words rather than keywords. While keywords are crucial for understanding, content words inflate user utterances. In an example from their study, a user would accentuate the words *Morning Meeting* in the utterance *Add an event called Morning Meeting* instead of *Add an event*, which is the keyword portion of the utterance. Recognition is hence not improving and “frustration leads to more frustration” [99, p. 5] and error spirals. Other strategies, such as simplification and supplying additional information rather confuse a CUI. In a conversation, CUIs are dependent on specific information at a certain point in time. Compared to humans, CUIs lack the flexibility and world knowledge to fall back on in case of errors.

Myers et al.’s [99] research shows that users struggle with building mental models of their CUIs’ capacities and lack the technical understanding of how a CUI handles errors. Although research paints a mixed picture, demographic factors seem to play a role in preferences for repair strategies. Ashktorab et al. [4] found preferences to depend on users’ orientation towards CUIs: a higher social orientation favored natural and more human-like repairs without error indications. A decrease in social orientation led to the preference of overall less natural but more effectual error handling strategies with fewer dialog turns. In general, simple repair strategies were preferred for simpler errors, while more complicated strategies were tolerated for more complicated errors. Mavrina et al. [89] on the other hand did not find a significant effect of internal reasons like attitude or age. Rather, they suspect situational aspects to impact strategy preferences.

2.4 Error Handling in Human-Human Interaction

Successful communication requires interlocutors to establish common ground. The process of mutually arriving *on the same page* was essentially coined by Clark and Brennan [31] and is referred to as grounding. In HCI, errors are oftentimes born from a lack of grounding [15, 99] due to “inadequate feedback and impoverished context” [15, p. 19]. Common ground is built and constantly updated by conversational partners. It is a collective and collaborative process culminating in mutual understanding. More specifically: In the so-called presentation phase, person A addresses Person B. In the subsequent acceptance phase, person B a) acknowledges person A has uttered something, and b) understands person A’s utterance. Grounding is accomplished if both phases are completed effectively, leading to a successful interaction. If an error occurs in one of the two phases, grounding has failed, and a suitable error handling strategy needs to be carried out [21]. To do so, human interlocutors have a rich arsenal of strategies at hand. These strategies span over verbal and non-verbal communication patterns, including e.g., eye-gaze or nodding. In fact, “specialized techniques have evolved for grounding different types of content … [and] grounding changes with the current purpose” [31, p. 136]. The selection of the correct error handling strategy thereby depends on the medium a conversation is held over and on strategy-specific costs. Building on Grice's [49] maxim of quantity and manner, Clark and Brennan [31] introduce the refined principle of Least Collaborative Effort. The principle describes that conversational partners "try to minimize their collaborative effort – the work that both do from the initiation of each contribution to its mutual acceptance" [31, p. 135]. This work can be more or less costly, meaning, it can be more or less elaborate. According to Clark and Brennan's principle of Least Collaborative Effort, conversational partners want to keep costs as low as possible. An exhaustive overview of all costs and medium-dependent constraints can be found in the researchers' paper [31], but some costs and constraints with relevance to HCI will be described in the following.

Constraints. Compared to HHI, HCI puts medium-dependent constraints forward which impede interactions between a given user and a CUI. Visibility is a key constraint. In most CUIs, non-verbal communication is not possible which complicates e.g., the coordination of dialog turns. Due to processing latencies, the constraints simultaneity, sequentiality, and revisability take effect, too. Conversational partners cannot send and receive utterances simultaneously and their turns can easily get out of sequence. Furthermore, revisability is restricted. Once an utterance is formulated, the possibility to correct it is limited.

Costs. Careful planning of utterances leads to formulation costs, whereby it "costs more to formulate perfect than imperfect utterances" [31, p. 142]. In case of an error and depending on the error handling strategy, formulation costs can vary. Repeating parts of an utterance is less costly than formulating a new utterance from scratch. Delay costs increase in case an utterance needs planning and editing. As errors accumulate, fault costs rise. Fault costs include an interlocutor's concern to appear "foolish, illiterate, or impolite" [31, p. 145]. Lastly, the more complex a repair and the more dialog turns it takes to resolve an error, the higher overall repair and turn-taking costs become.

In his work on HHI error handling strategies, Skantze [130] found that conversational partners seldomly signal errors directly (as in *I'm sorry, I didn't catch that*). Rather, study participants tried to advance dialogs by continuing descriptions, asking task-related questions, and making assumptions regarding their conversational partners' intent. While signaling errors had a negative impact on study participants' experience of task success, asking task-related questions led to a fast error recovery. Referring back to Clark and Brennan [31], faster recovery means complying with the principle of Least Collaborative Effort, which reduces costs. While users were faced with numerous non-understandings in Skantze's [130] study – a word error rate of 42% is reported – study participants reported that they had almost always been understood. Schegloff et al. [119] adds to the research around HHI error handling strategies by introducing the terms self- and other-repair. Self-repair originates from the side of the conversational partner who produced an error. Other-repair on the other hand is conducted by a given listener. Schegloff found self-repair to be interlocutors' preferred form of repair. Again, Clark and Brennan's [31] cost model as well as their principle of Least Collaborative Effort can be consulted to explain this preference. Fault costs as well as turn-taking costs increase with other-repair and make this strategy more costly. Self-initiated repairs can often be carried out in the erroneous dialog turn. Contrarily, other-repair may need multiple turns to solve an error. In these cases, a dialog turn is blocked for error recovery, thereby not furthering the conversation. Dialing back to HCI, Cuadra et al. [35] find benefits and preferences for self-repair to also manifest in interactions with CUIs. CUIs engaging in self-repair were thereby assessed to be more intelligent.

3 METHOD

To gain insights into preferences regarding error handling strategies for CUIs, HHI informed error handling strategies were developed and evaluated in a within-subjects design in a preliminary crowdsourcing and a WoZ-led driving simulator study.

3.1 Research Design

A mixed methods design with quantitative and qualitative measurements was designed to answer this paper's research questions and hypotheses. To test the influence of the independent variable *error handling strategy* on the dependent variable *strategy ratings*, the research design comprised the execution of a preliminary crowdsourcing study, a driving simulator study, and a structured post-study interview.

Users of CUIs in the car were defined as the population of interest. To draw a sample of this population, German native speakers with experience with in-car CUIs and a valid driver's license were recruited for the preliminary crowdsourcing and the driving simulator study. A background in CUI design and implementation disqualified candidates from participating. Crowdsourcing study participants were recruited via the platform defined.ai. Defined.ai [36] is committed to ethical crowd work and reimburses workers with at least the minimum wage per locale. Internal BMW employees were recruited for the driving simulator study by means of an internal company platform. Their attendance was not compensated monetarily as they could participate during regular working hours.

Data was collected in form of quantitative observations with a Likert scale survey for the crowdsourcing study, the validated UEQ+ questionnaire [67] for the simulator study, and a structured post-interview in form of subjective rankings. The calculation of costs was facilitated through anonymized audio recordings documented in the driving simulator study and the subsequent analysis of transcriptions via WebMAUS Basic [66]. Data analysis methods are described in detail in the following chapters. All study participants were informed about their right to withdraw from the studies at any time and without consequences. For the driving simulator study, participants were informed about the driving scenario upfront. It was made transparent that they would not encounter invasive, dangerous, stressful, or extreme driving situations. Participants were informed that their interactions were recorded via audio and had the possibility to withdraw their consent. Furthermore, they were made aware of the potential occurrence of motion sickness and respective safety protocols for terminating the simulation by themselves via an emergency button. All study participants gave written approval. Upon completion of the studies, participants were given the opportunity to ask questions and leave remarks regarding their experiences. Moreover, a short debriefing informed them about their right to object to the processing of their data and request data deletion.

3.2 Development of Error Handling Strategies

Based on the theoretical framework outlined in the chapters above, three human-centred error handling strategies were developed for the present study. Subsequently, these strategies were compared to a common HCI baseline strategy, namely signaling non-understanding, and asking users to repeat themselves. Error handling strategies were systematically varied regarding error confession (confession vs no confession), question type (open vs closed question), and adequacy of suggestion (correct vs incorrect suggestion). Examples for strategies are given for a calling use case where the name of the contact was not recognized (Possible user utterance: *Please call Kate*)[20].

Baseline Condition – (*baseline*). Example: *I'm sorry, I didn't catch that. Could you please repeat that?* In this *baseline* condition, non-understanding is signaled, and users are asked to repeat their utterance. Signaling uncertainty and indicating errors can help users to update their mental

[20] The study was conducted in German; English translations are provided by the author.

model of a CUI, helping them to explore and learn system boundaries [87, 99]. On the other hand, signaling non-understanding can lead to error spirals. Users are found to repeat utterances verbatim instead of rephrasing them. Alternatively, they start hyper-articulating which does not facilitate NLU [86, 99, 130]. The concomitant costs for this strategy are potentially high. Although repeating utterances is less costly than formulating new ones, delay costs can increase if users plan a refined and edited version of their initial utterance [31]. The number of turns needed to repair a dialog depends on the reason for an error: if an utterance was not recognized due to e.g., ambient noise, repeating it can lead to a successful dialog. However, if a user's choice of words and/or system boundaries led to the error, repeating the initial utterance will lead to an error spiral.

Strategy 1 – Task-related questions (*task-related*). Example: *Who do you want to call? task-related* questions attempt to further a conversation and work towards the joint goal of completing a task. As proposed as a best practice in current research [4, 130, 147] and common in HHI error handling strategies [86, 130], errors are not indicated. Rather, users are asked concrete and *task-related* questions. The costs for this strategy are as follows: formulation costs are low as the required user answer can consist of only one word and the formulation of a "perfect utterance" [31, p. 142] is not necessary. Due to the question type being an open question, the number of turns needed to repair the dialog is most likely limited to one turn.

Strategy 2 – Anticipatory error handling (*anticipatory*). Example: *Do you want to call someone?* As in strategy 1, an *anticipatory* error handling strategy does not directly indicate an error. Formulation costs are low as users do not need to form a complete sentence. Depending on the number of turns used for the repair, turn-taking costs are potentially higher than in Strategy 1. Due to the question being a closed question, dialog turns can vary between one or two further turns. Users can opt to repair the dialog in one turn (as in e.g., *Yes, Kate*) or in two turns, requiring the assistant to also prompt again (as in e.g., user: *Please call Kate* – CUI: *Do you want to call someone?* – user: *Yes* – CUI: *Who do you want to call?* – user: *Kate*).

Strategy 3 – Query (*query*). Example: *Did I understand correctly that you want to call someone?* In this strategy, the CUI indicates uncertainty and provides a task-related closed question furthering the dialog. The error is not addressed as clearly as in the *baseline* condition. Due to the closed question type, the number of turns to repair the dialog can amount to two turns (see example in Strategy 2). Again, formulation costs depend on the number of turns, but are potentially higher than for Strategy 1.

The strategies above assume the CUI has correctly understood a user intent (*calling someone*) and can hence provide users with correct task-related questions. If an intent is understood incorrectly or if the confidence level of understanding is very low, the strategies can still be adopted. Acceptance and strategy preferences may deviate in these cases though. To account for this, each error handling strategy (except for *baseline*) was designed for adequacy of suggestion, meaning a) the intent was recognized correctly, and a correct suggestion is made, and b) the intent was recognized incorrectly, and an incorrect suggestion is made. All interactions are listed in detail in Table 1.

Table 1: Overview over Interactions

Error Handling Strategy	**Exemplary User Utterance**	**Error Handling Prompt**	**Question Type**	**Error Confession**	**Adequacy of Suggestion**
baseline	Connect my smart phone	I'm sorry, I didn't catch that. Could you please repeat that?	closed	yes	not applicable
task-related	Look for a gas station	What are you looking for?	open	no	yes
anticipatory	Call Kate	Do you want to call someone?	closed	no	yes
query	Activate the seat heating	Did I understand correctly that you want to activate something?	closed	yes	yes
task-related	Navigate to Klenzestraße	What do you want to open?	open	no	no
anticipatory	Play 99 ballons	Do you want to start a navigation?	closed	no	no
query	Open the window	Did I understand correctly that you want to listen to something?	closed	yes	no

3.3 Cost Calculations

To determine costs per strategy, the average number of turns and the average number of words can be calculated using the following formula:

$$costs = \emptyset \text{ number of words} + \emptyset \text{ number of turns}$$

Exemplarily: the average number of turns in strategy X amounts to 2, and the average number of words in strategy X amounts to 5. Costs for this strategy will be added up as 5+2=7. Based on the considerations around turn-taking and formulation costs established in chapter 3.2, the order of costs per strategy is hypothesized as follows (from lowest to highest costs): Task-related questions (*task-related*), Anticipatory Error Handling (*anticipatory*), Query (*query*), and Baseline (*baseline*).

To account for potential differences in dialog “wordiness” due to user-specific use of language and varying use cases (e.g., calling somebody vs searching for a gas station), another insightful metric can be calculated. The proposed *repair ratio* describes the relation between the words used in the initial user utterance and the number of words used to repair a dialog. The repair ratio can be calculated by dividing the number of words needed to repair a dialog by the number of words forming the initial utterance, times 100.

$$repair\ ratio = \frac{\text{number of words used to repair}}{\text{number of words used for initial utterance}} * 100$$

Exemplarily: the number of repair words in a dialog amounts to 5, and the number of words in the initial utterance amounts to 10. The repair ratio is 5/10*100=50%. This metric allows the clustering of formulation costs in low (repair ratio ≤50%), high (repair ratio >50%<100%), and very high (repair ratio ≥100%). Combined with the number of turns, both parameters will be used for the statistical analysis of costs.

3.4 Preliminary Study

A preliminary crowdsourcing study was conducted to ensure the comparability of error handling strategies in terms of naturalness. The synthetic text-to-speech (TTS) voice did not change between strategies. Nonetheless, a preliminary study controls it as a potential confounding factor. All dialogs planned to be used in the simulator study were recorded as mock-up dialogs between the author and the used CUI. These dialogs were subsequently presented to study participants via audio files in a within-subjects crowdsourcing study. 200 study participants were asked to rate the TTS voice presenting the strategies on a seven-level Likert scale, ranging from very natural to very unnatural. A full overview over study dialogs can be found in Appendix A.

Due to the categorical nature of the data obtained in the crowdsourcing study, Chi-Square tests were calculated in R [116]. With χ^2(36, N=200)=23.96, p=.94, no significant differences were found between the strategies’ naturalness ratings. Hence, no influence of the TTS voice was assumed, and all strategies were adopted for the driving simulator study.

3.5 Driving Simulator Study

As the present paper aims at examining error handling strategies in in-car CUIs, the concomitant study was performed in a high-fidelity driving simulator. Driving simulators offer a controlled and reproducible study environment and are “predictive for on-the-road driving performance” [7, 143, 144]. By accounting for external factors such as traffic, traffic noise, and the execution of the primary driving task, driving simulator studies provide the possibility to examine HCI interactions in a realistically simulated as well as standardized traffic situation, without compromising consistent study conditions and safety of study participants. Driving simulator studies bear the risk of motion sickness, which can manifest as nausea, dizziness, or headaches. However, the careful selection of a suitable driving scenario with a low degree of turn-taking, refraining from infrequent stops, and driving at lower speeds can mitigate this risk [58]. Depending on the degree of simulator fidelity, simulator studies can lead to unrealistic driving behavior or generate a false sense of safety that is argued to translate back to actual in-world driving [143, 144]. Hence, the present driving scenario was implemented in a high-fidelity simulator and the driving scenario was designed with Hwangbo et al.'s [58] recommendations in mind. The usage of a high-fidelity simulator with an extensive field of view as well as a car mock-up lead to high immersion in the driving task. This immersion allows the applicability of results and the derivation of concrete design guidelines to the real world [88]. However, even a high-fidelity simulator cannot completely divert attention from the lab setting - if only because study participants are repeatedly asked for evaluations.

3.5.1 Driving Simulator and Driving Scenario

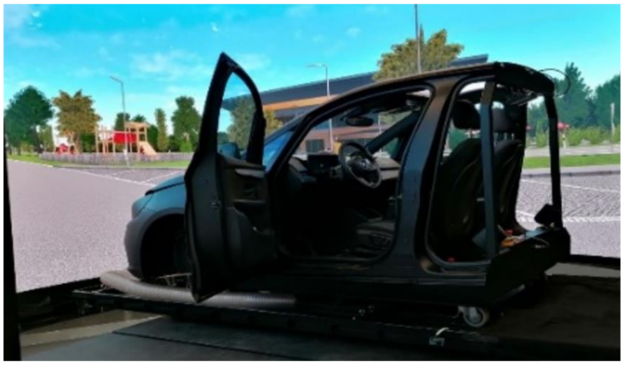


Figure 1: Driving Simulator Study Setup

The driving simulator study was held in a static driving simulator with a 180° screen and a stationary vehicle mock-up (see Figure 1). The mock-up was equipped with a CUI, which was connected to an operator desk. Speech interactions were executable by the experimenter in a Wizard of Oz manner. A highway setting with overall low traffic density was chosen as driving scenario for the present study. Driving simulator studies continually report low secondary task engagement and poor user evaluations due to primary task overload [14, 63, 106]. As the study's focus lay on the secondary task, namely the evaluation of interactions, the primary driving task was designed with these results in mind. To find a trade-off between simulating a realistic driving scenario without overloading users with the primary driving task, the driving scenario was designed plain and straightforward. Study participants were asked to follow a lead vehicle on the right highway lane with a speed of 100km/h. Conditions did not vary between study participants and throughout the experiment drive.

To test the cognitive load inflicted by the study, a preliminary trial run was conducted with five study participants. Cognitive load was self-reported by study participants by means of the Driver Activity Load Index (DALI) [107]. The DALI reports cognitive load on a scale of 1 (low) to 100 (high). Study participants reported an overall workload of 52/100 (n=5, sd=11.66), and an interference between the primary and the secondary task of 50/100 (n=5, sd=4.71). Both can be categorized as a medium workload. The obtained insights led to the decision to maintain the driving scenario described above and to not increase the complexity of the primary driving task.

3.5.2 Questionnaires and Measurement Tools

Prior to the driving task, study participants answered demographic questions around age, gender, and experience with in-car CUIs. Furthermore, a questionnaire surveyed if study participants' motivation to use CUIs was hedonic, social, or utilitarian [92]. As attitudes and motivations to use CUIs can vary, this measurement accounts for potential individual differences in interactions with CUIs.

Subsequently, study participants were familiarized with the study use cases and the structure of interactions. In order not to prime initial user commands, use cases were presented as pictograms. Before the experiment started, the experimenter asked participants to verbalize all commands to ensure their comprehensibility. In general, all commands were easily understood and formulated as expected. In case a formulation differed from the target utterance, study participants were asked for alternative formulations. Once the intended formulation was used, the experimenter affirmed it. A list with the graphical representation of use cases remained in the study participants' field of view throughout the entire study drive. User utterances during the study did not differ notably from the pre-study test run, although few study participants used synonyms as in e.g., pre-study utterance: *start the seat heating* vs study utterance: *activate the seat heating*. The pictograms are displayed in Figure 2 (see Table 1 for an overview over intended user utterances).

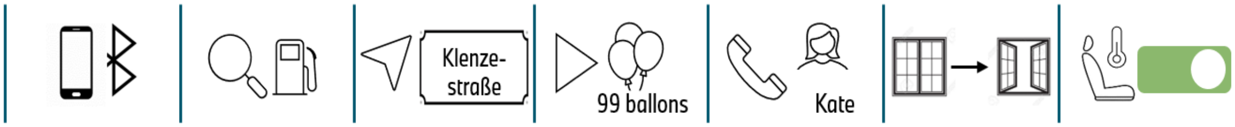


Figure 2: Pictograms Displaying Study Use Cases

Strategy preferences were measured twice. First, quantitatively by means of the UEQ+ [67], and secondly qualitatively after the drive. The UEQ+ is a repeatedly validated modular questionnaire containing various scales, which measure the user experience of interactive systems [67, 125]. The questionnaire provides the items *response quality*, *response behavior*, and *comprehensibility*, which are specifically tailored to evaluate interactions with CUIs [67]. As the present study focusses on the structure of dialogs and the quality of CUI answers, the items *response behavior* and *response quality* were selected from the UEQ+. However, the item *comprehensibility* was omitted in order not to shift study participants' attention to the erroneous nature of the dialogs. *Response behavior* as well as *response quality* were queried after each interaction. Both scales contain four contrasting adjective pairs, which can be evaluated on seven-level Likert scales. For the second qualitative evaluation, study participants were presented with all error handling strategies in written form and were asked to subjectively rank strategies from best to worst.

At the end of each drive, the perceived overall workload of the driving simulator study was self-reported with the help of the DALI questionnaire [107]. Since strategy preferences were the primary focus of research, we refrained from measuring the cognitive load after each interaction. As interactions were designed consistently across error handling strategies (see chapter 3.6), it can be assumed that they do not differ fine-grained in terms of cognitive load. Anecdotally, study participants did not indicate problems with the driving task (such as e.g., problems with lane keeping), nor did the experimenter report deviating driving behavior due to cognitive overload.

3.5.3 Execution of the Driving Simulator Study

Upon arrival, subjects gave informed consent and answered demographic questions (see chapter 3.5.2). A three-minute familiarization drive preceded the actual experiment and gave participants the possibility to get used to the vehicle and the simulated route. As the study was set up in a within-subjects design, all study participants experienced and subsequently rated the same seven erroneous dialogs (see chapter 3.6). Upon completion of an interaction, participants were presented with the standardized UEQ+ questionnaire [67] and were asked to rate the experienced dialog on the scales *response behavior* and *response quality*. After the last interaction, study participants' cognitive load was measured by means of the DALI questionnaire [107]. In the post-study interview, study participants were presented with all error handling strategies in written form and asked to rank strategies from best to worst, to obtain qualitative insights into preferences for strategies. In total, the study was completed in approximately 45 minutes, whereby the drive took around 20 minutes per participant.

3.6 Interactions

In total, seven error handling strategies were presented to study participants in the driving simulator study. Strategies were systematically varied regarding error confession (confession vs no confession), question type (open vs closed question), and adequacy of suggestion (correct vs incorrect suggestion).

All interactions were triggered from the study participants' side, following the pictograms detailed in chapter 3.5.2. Study participants were asked to give a verbal command to the CUI and follow the ensuing dialog naturally and intuitively. Commands were queried in randomized order to avoid sequence effects. All interactions followed the same pattern, which is displayed in Figure 3.

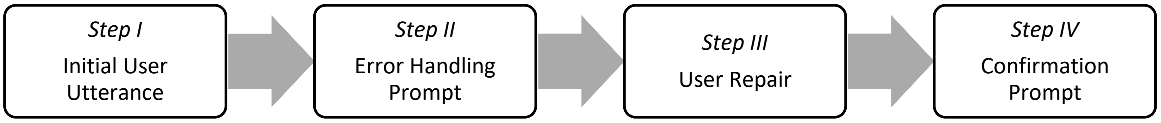


Figure 3: Interaction Structure

After every initial user utterance (step 1), the experimenter triggered an error in a Wizard of Oz manner (step 2) which subsequently had to be repaired by study participants (step 3). Upon completion of the repair, the experimenter went on to a confirmation prompt (step 4). In theory and depending on the user repair, study participants could complete all interactions within two dialog turns.

To standardize interactions between study participants, no further errors were introduced by the experimenter after step 3 of the interaction process. Although this procedure restricts the overall number of dialog turns and the position of errors, it is crucial to generate standardized dialogs and deduce generalizable results. If the concept of costs is applicable to HCI, their impact should be detectable independent of the number of dialog steps. Dialogs were scanned for errors in step 1) of interactions (study participants' initial utterances). A higher tolerance for CUI errors is highly likely in case users produce an error. Furthermore, initial as well as repair utterances were transcribed and utilized to calculate word costs per strategy. A large body of research suggests that users tailor their error handling strategy to the perceived type of error [46, 64, 97, 99]. Study participants' repair strategies may therefore vary regarding word costs if they assume an error originated from their side. As such deviations potentially skew strategy costs, the data set was cleared of these interactions.

4 RESULTS

n=48 study participants experienced and rated seven erroneous dialogs with an in-car CUI. To answer this paper's research questions and hypotheses, repeated-measures ANOVAs were calculated, observing error handling strategies as a within factor. Furthermore, qualitative strategy rankings were evaluated by means of Chi-Square tests. As described in detail in chapter 3.6, all interactions followed the same structure, namely 1) initial user utterance, 2) error handling strategy, 3) user repair, and 4) CUI confirmation. As discussed, dialogs were scanned for errors in step 1) of interactions (study participants' initial utterances) and subsequently cleared of these instances. From an initial 336 interactions, 34 interactions produced an error in the initial user utterance, leaving 302 ratings by 48 study participants for the statistical analysis.

4.1 Study Participants

A total of n=48 participants (68.75% male and 31.25% female) participated in the driving simulator study. The average age of participants was 36.8 years (sd=10.61). All but 8.33% of study participants reported experience with in-car CUIs, with 12.5% indicating to use their in-car CUI during every drive, 18.75% during every second drive, and 43.75% to use it less frequently. Study participants' attitude towards CUIs was measured following McLean & Osei-Frimpong [92] and revealed 47.92% utilitarian attitudes, 41.67% hedonic attitudes, 2.10% social attitudes, and 8.33% of study participants with a mixed attitude. Users with utilitarian attitudes use CUIs from a mostly task-oriented perspective, whereas a hedonic use is characterized by seeking joy and pleasure. Socially driven users on the other hand interact with CUIs for social reasons [92]. Study participants' workload was measured by means of the DALI questionnaire at the end of the simulated drive [107]. Overall, DALI ratings for the driving simulator study indicated a low workload of 39/100 (n=48, sd=20.0) as well as a low interference between the primary and the secondary task of 40/100 (n=48, sd=3.73). More specifically, 85.4% of study participants indicated that their experienced workload was very low, low, or medium, with only 14.58% of study participants indicating a high workload. None of the participants declared a very high workload.

4.2 Statistical Analysis of Costs

To determine costs per strategy, the average number of dialog turns, and the average number of repair words were obtained and summed up for each strategy. Dialog turns were thereby calculated manually, while the number of words was calculated on the basis of transcripts of study dialogs gathered from WebMAUS Basic [66]. Costs were computed per participant and strategy, using the formula specified in chapter 3.3:

$$costs = \emptyset \text{ number of words} + \emptyset \text{ number of turns}$$

Figures 4 and 5 provide an overview over all tested error handling strategies, their concomitant word, and turn-taking costs as well as the sum of costs. Bars marked in shades of green thereby display strategies with correct suggestions, while bars marked in red show costs of strategies with incorrect suggestions. The *baseline* condition is displayed in both bar charts.

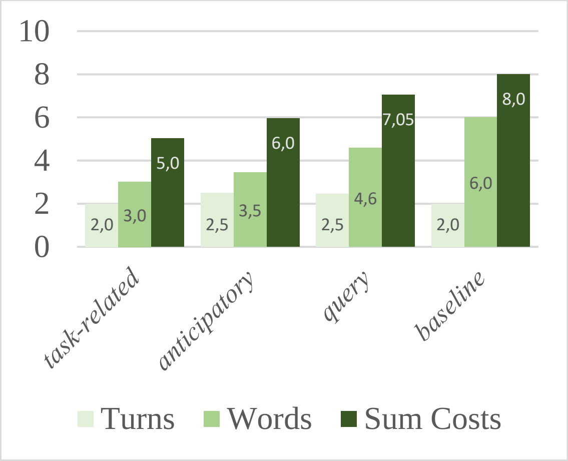


Figure 4: Turns, Words, and Total Costs for Error Handling Strategies with Correct Suggestions

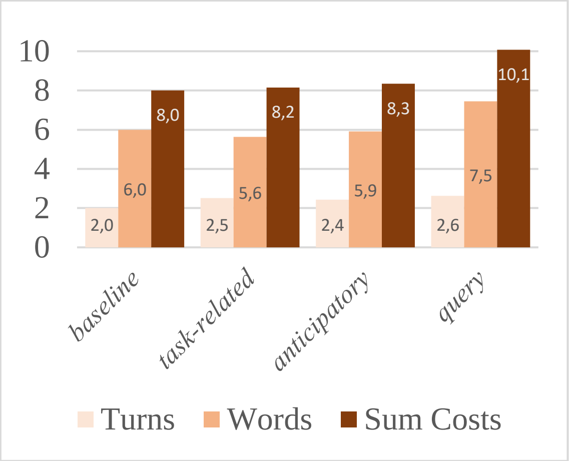


Figure 5: Turns, Words, and Total Costs for Error Handling Strategies with Incorrect Suggestions

Error handling strategies were found to differ significantly regarding their concomitant repair costs. With F(3, 155)=23.14, p<.0001, HHI-informed error handling strategies (*task-related*, *anticipatory*, and *query*) were associated with significantly lower repair costs than the HCI *baseline* condition. While costs were found to differ significantly between strategies, study participants' UEQ+ ratings did not mirror this result. With F(1, 273)=2.2, p=.14, the UEQ+ items *response behavior* and *response quality* did not capture a significant difference between error handling strategies in terms of their user experience. This result is displayed in Figure 6.

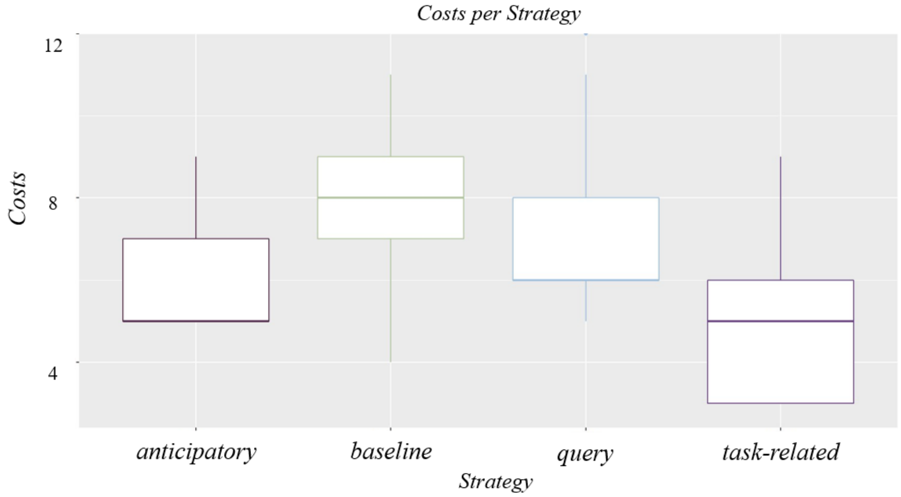


Figure 6: Costs per Strategy

4.2.1 Preferences for Error Handling Strategies with Correct Suggestions

As error handling strategies were varied systematically regarding the adequacy of the proposed suggestion, results are reported separately for strategies making a correct suggestion, and strategies making an incorrect suggestion. The importance of the adequacy of suggestion is reflected in study participants' UEQ+ ratings. With F(2, 294)=11.5, p<.0001, strategies delivering a correct suggestion were preferred over strategies delivering an incorrect suggestion. The importance of correct suggestions is furthermore reflected in repair costs. Strategies with incorrect suggestions led to a significantly higher number of dialog turns needed to repair an interaction, than strategies with correct suggestions: F(2, 272)=17.68, p<.0001. The same effect was found for the number of words needed to repair a dialog. With F(2, 272)=59.29, p<.0001, strategies with incorrect suggestions needed more words to be repaired than strategies with correct suggestions. Chi-Square tests were calculated to analyses study participants' qualitative strategy rankings. With χ^2(9, N=48)=78.3, p<.0001, they revealed a significant relationship between error handling strategies and study participants' qualitative rankings. With Cohen's w for cross tables producing 0.64, this effect is classified as large. As can be seen in Figure 7, the least costly strategy *task-related* was ranked as preferred error handling strategy on Rank 1, followed by the low-cost strategy *anticipatory* on rank 2, the more expensive strategy *query* on rank 3, and the most expensive strategy *baseline* on rank 4.

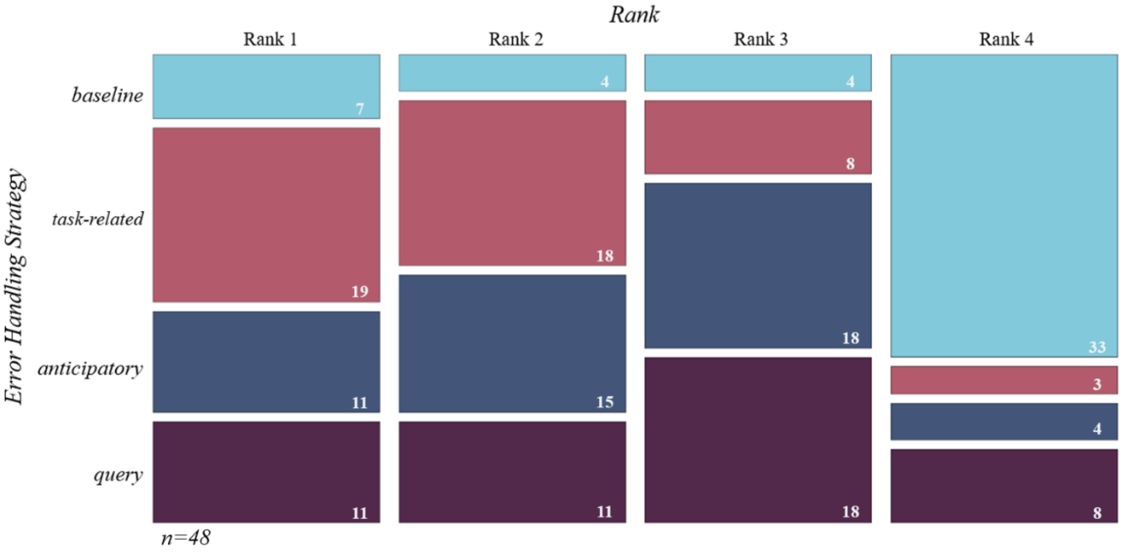


Figure 7: Qualitative Ranking of Error Handling Strategies with Correct Suggestions

4.2.2 Preference for Error Handling Strategies with Incorrect Suggestions

Within error handling strategies with incorrect suggestions, Chi-Square tests also detected a significant relationship between error handling strategies and the qualitative rankings of study participants: $\chi 2$(9, N=48)= 85.3, p<.0001. With 0.67, Cohen's w for cross tables detected a large effect. Figure 8 shows baseline to be the preferred error handling strategy in case an incorrect suggestion was made by the CUI. This strategy is followed by task-related on Rank 2, anticipatory on Rank 3, and query on Rank 4. As for strategies with correct suggestions, this ranking mirrors the cost structure of strategies: baseline as least costly strategy is preferred by study participants, while the most cost-intensive strategy – query – is ranked last.

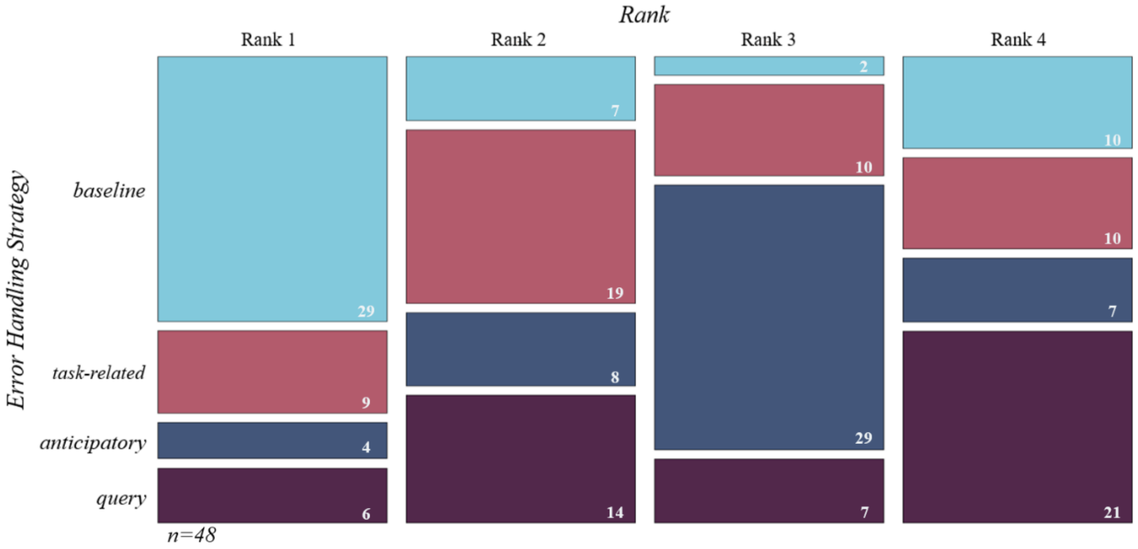


Figure 8: Qualitative Ranking of Error Handling Strategies with Incorrect Suggestions

4.3 Further Analysis of Error Handling Strategies

Next to the evaluation of UEQ+ ratings, Pearson correlation coefficients were calculated to analyse study participants' strategy rankings obtained in the post-study interview. Results shed light on the effects of different strategy components as well as on repair costs. To account for potential further influence factors on study participants' evaluations, the effect of demographic factors, the question type, as well as the confession of errors was analyzed. Furthermore, the length of dialogs, the repair ratio, the number of hesitations, as well as the number of full dialog resets[21] per strategy were collected and transcribed. Results are reported for UEQ+ ratings, costs, as well as correlations where appropriate. At the end of the chapter, Table 2 provides a conclusive overview over all influence factors on error handling strategy preferences.

4.3.1 Demographic Factors

No influence was found for study participants age (p=.83) and frequency of in-car CUI use (p=.23). Gender of participants significantly influenced UEQ+ ratings, in that female study participants generally rated interactions better than male study participants, F(1, 269)=14.78, p=.0002. However, there was no interaction between gender and strategy preferences (p=.94). Furthermore, the attitude towards CUIs did not influence strategy ratings (p=.05), which is substantiated as no

[21] Full dialog resets meaning a dialog restart with a fallback on generic error handling in form of "Okay, how may I help you?"

correlation between qualitative strategy rankings and demographic factors was detected: *r*(190)=.0.

4.3.2 Question Type

While the question type did not influence study participants' ratings of interactions on the UEQ+, it significantly impacted the number of turns needed to repair a dialog. Compared to open questions, closed questions (as in *anticipatory* & *query*) led to an increase in the number of dialog turns, F(1, 334)=5.62, p=.0018. Furthermore, the question type significantly influenced the number of words needed to repair a dialog, F(2, 280)=5.44, p=.005, in that closed questions led to fewer words per turn. A medium negative correlation was found between the question type and strategy rankings for error handling strategies with correct suggestions, *r*(190)=-.31, p<.0001.

4.3.3 Error Confession

Errors were confessed in the *baseline* condition as well as for *query*. No influence of error confession was found for study participants' ratings of dialogs on the UEQ+ (p=.27). Furthermore, error confessions did not influence the number of turns (p=.3), or words (p=.19) needed to repair a dialog. There was, however, a medium positive correlation between error confessions and strategy rankings for strategies with correct suggestions, *r*(190)=.35, p<.0001.

4.3.4 Dialog Length

While no influence on UEQ+ evaluations was found for the length of dialogs (p=.49), some strategies led to significantly longer dialog durations than others, F(6, 276)=66.28, p<.0001. *baseline* was found to be significantly longer than *task-related*, *anticipatory* (p<.0001), and *query* (p=.005). Furthermore, *task-related*, and *anticipatory* were significantly shorter than *query* (p<.0001). Length thereby equally depended on the number of turns (p=.0002) and the number of words (p<.0001) used to repair a dialog. A small positive correlation was found between length of dialogs and strategy rankings for strategies with correct suggestions, *r*(190)=.18, p=.013.

4.3.5 Repair Ratio

The repair ratio describes the ratio of number of words used for an initial user utterance and the number of words needed to repair a dialog. In case of incorrect suggestions, the repair ratio amounted to 105%, meaning the number of words needed to repair an interaction was 105% higher than the number of words used in the initial utterance (e.g., initial utterance: 5 words, repair: 10.25 words). For interactions with correct suggestions, this number was as low as 60.6%, leading to a significant difference in the repair ratio due to the adequacy of suggestion: F(2, 272)=66.84, p<.0001. Compared to all other strategies, *task-related* was thereby found to have the lowest ratio with 44.4% (p≤.03). As such, a small positive correlation was found between repair ratio and strategy rankings, *r*(190)=.18, p=.013.

4.3.6 Number of Hesitations

Hesitations (as in *mmh*) were recorded in the transcription process but were not found to depend on the used error handling strategy (p=.83) or influence UEQ+ ratings (p=.16). Furthermore, strategy rankings and number of hesitations did not correlate, *r*(190)=.0.

4.3.7 Full Dialog Resets

The number of full resets necessary when repairing a dialog was found to depend on the adequacy of suggestion. While no full resets were needed for strategies providing correct suggestions, incorrect suggestions led to up to almost half of all interactions needing full resets. More specifically, *task-related* led to 4.9%, *query* to 41.5%, and *anticipatory* led to 45.2% of interactions needing full dialog resets. Due to their inherent structure, all *baseline* dialogs can effectively be considered full resets. Resets caused a significant increase in dialog turns, F(1, 273)=107.41, p<.0001. Still, no correlation between strategy rankings and dialog resets (*r*(190)=.0) as well as no effect of dialog resets on UEQ+ ratings was found (p=.72).

The following table summarizes this chapter's findings and gives a conclusive overview over influence factors playing a role in evaluations and rankings of error handling strategies.

Table 2: Overview over Influence Factors on Error Handling Strategies

	UEQ+	**Costs**	**Correlations**		**UEQ+**	**Costs**	**Correlations**
adequacy of suggestion	sig	sig	sig	*dialog length*	ns	sig	sig
demographic factors	ns	ns	ns	*repair ratio*	ns	sig	sig
question type	ns	sig	sig	*hesitations*	ns	ns	ns
confession of errors	ns	ns	sig	*dialog resets*	ns	sig	ns

5 DISCUSSION

Initially, the role of the UEQ+ as well as qualitative rankings need to be discussed. Except for adequacy of suggestion, no significant differences in strategy ratings were found on the UEQ+ scales *response behavior* and *response quality*. A qualitative strategy ranking on the other hand did reveal significant preferences for strategies. Moreover, qualitative rankings significantly coincided with repair costs, thereby capturing user preferences in a more granular way. The discrepancy between UEQ+ ratings and qualitative rankings suggests that the UEQ+ can only partly capture the user experience of erroneous dialogs in an in-car environment. Still, response behavior and response quality are important levels to consider when designing error handling strategies. This finding underlines the need for fine-grained measurement tools to expand evaluation possibilities for CUIs. Future work could undertake further studies on whether the UEQ+ or different UEQ+ items can be validated explicitly for error cases. For instance, studies could examine whether the UEQ+ item *efficiency* correlates with cost calculations.

5.1 Discussion of Costs

HHI-informed error handling strategies were found to be significantly less costly than a compared classic HCI error handling strategy. This means that H1, namely *HHI informed error handling strategies lead to lower repair costs than a HCI error handling strategy*, can be supported. The principle of Least Collaborative Effort states that conversational partners intend to keep costs as low as possible when repairing erroneous conversations. Following this principle, human-centred error handling strategies were found to be significantly more cost-efficient than a compared common HCI error handling strategy.

Error handling strategies with low costs were thereby significantly preferred by study participants in a qualitative ranking. For error handling strategies making correct suggestions, *task-related* error handling, the strategy associated with the lowest repair costs, was judged to be study participants' preferred error handling strategy. Trying to further a conversation by using *task-related* or *anticipatory* questions is a mechanism used by human interlocutors [130] and advisable as long as the context is understood correctly. This is in line with findings by Zargham et al. [147] who report positive outcomes for anticipatory error handling strategies only in case the proposed suggestion matches the user intent. The strategy of asking correct cost-efficient *task-related* questions can only be administered in case the current user context is known and an adequate suggestion to solve an error can be made. In case an incorrect suggestion was made by the CUI, repair costs for HHI-driven strategies rose. With the *baseline* strategy being ranked highest in this condition, study participants again preferred the least cost-intensive error handling strategy to repair their dialogs. Nonetheless, asking *task-related* questions was ranked closely behind the *baseline* condition. Even in case an incorrect suggestion was made by the CUI, repair costs for this strategy remained low. With this finding, RQ2 is answered and H2 – *The error handling strategy associated with the lowest repair costs is preferred by study participants* – is supported. In conclusion, costs were found to be a significant indicator for error handling strategy preferences. As such, they are an easily applicable and straightforward measurement tool when designing conversational repairs in HCI.

5.2 Discussion of further Strategy Analyses

Additional systematic insights into error handling preferences, costs, as well as cost drivers were gained by observing how conversations unfolded during the driving simulator study.

Partly in line with current research, demographic factors were not found to influence strategy preferences, although it needs mentioning that gender as well as study participants' attitudes towards CUIs were not evenly distributed in our sample. As in Mavrina et al.'s [89] study, age, gender, experience with in-car CUIs, and attitude towards CUIs did not play a role in study participants' strategy rankings. Although Ashktorab et al. [4] did find an influence of attitude on error handling preferences, the present study cannot mirror these results.

Considerations around costs were found to provide a more conclusive explanation for strategy preferences. Next to incorrect suggestions, the question type led to an increase in repair costs. Coherently, increased word and turn-taking costs led to significantly longer overall dialog lengths. *task-related* as well as *anticipatory* were thereby repaired in a significantly shorter amount of time than *query* and *baseline*. Choosing a closed question type as in *anticipatory* and *query* led to a significant increase in the number of turns study participants needed to repair a dialog. At the same time, a higher number of turns led to a significant decrease in the number of repair words per turn. Study participants repairing dialogs in more turns potentially perceived the CUI as less intelligent and hence used fewer repair words and stepwise repairs. This is supported by Myers et al.'s [99] findings that a more intelligent CUI will be provided with additional information to repair a dialog.

In line with related research, error confessions negatively impacted strategy ratings in case of correct suggestions. At the same time, error confession was not found to influence costs as it affected neither the number of turns, nor the number of words needed to repair a dialog. Still, *query*, and *baseline* – the two strategies confessing errors – were ranked behind *task-related* and *anticipatory* which both do not explicitly indicate errors. While the error confessing *baseline* strategy was preferred in case the CUI made an incorrect suggestion, *query* – which also indicates an error – was ranked lowest in these cases. Hence, the conclusion can be drawn that the confession of errors is not advisable, even if only incorrect suggestions can be made. This adds to findings from Ashktorab et al. [4] who discovered lower CUI likability and lower perceived CUI intelligence in case errors were confessed in task-related use cases. The present study seconds this detrimental effect of error confessions. With this finding, our research is in line with observations of error handling strategies in HHI, where non-understandings are seldomly signaled directly [130].

5.3 Limitations and Future Work

While a balanced gender ratio was pursued for the present study, two thirds of study participants identified as male. Although no effect of gender on strategy preferences was found, this imbalance needs reporting. Furthermore, the study was conducted in German and findings may be language and/or culture dependent.

Analyzed study use cases were without exception task-oriented and specifically tailored to suit an in-car setting. Error handling strategies in security-relevant cases or stressful situations as well as outside the car may require their own set of design rules. Future work can concentrate on further evaluating the suitability of costs and the principle of Least Collaborative Effort on regular, task oriented HCI dialogs of various domains. Furthermore, an additional weighing of word and turn-taking costs could prove beneficial for different languages (with higher/lower degrees of wordiness and higher/lower degrees of syntactical complexity) and NLU models or could even be user-specific.

The present study did not take context variables like e.g., eye-tracking or gesture recognition into account. These metrics can already be employed in the vehicle though. As human interlocutors repair errors multimodally, this research direction is an intriguing additional starting point for facilitating error handling in HCI even further. Although the UEQ+ seconds the importance of adequacy of suggestion, it did not detect further differences in the user experience of erroneous dialogs. Cost considerations as well as qualitative rankings on the other hand demonstrated significant preferences for human-centred error handling strategies.

As visually or hearing-impaired users are oftentimes restricted from driving, this user group was not examined in the present study. However, a growing body of research finds large differences in these or towards CUIs. For instance, a study found that blind users can handle a much higher degree of linguistic complexity than their sighted peers [13]. Although high pitched TTS voices and CUIs in environments with background noise are challenging for users with hearing impairments, CUIs pose a fruitful opportunity for these users to handle complex tasks [113]. However, especially in error cases, hearing impaired users were found to prefer visual error handling strategies [10]. The applicability of costs could potentially be leveraged for user groups with e.g., visual or hearing impairments. However, it is mandatory to also consider other parameters such as multimodality, TTS pitch and speed, or microphone opening times.

The data set was cleared of interactions with an actual error in study participants' initial utterances to prevent a possible effect on UEQ+ ratings. As users are furthermore known to adapt their error handling strategies to the source of an error, an impact of erroneous user utterances on repair costs was suspected. Although a comparison of ratings in these and error free cases is compelling, the low number of erroneous initial utterances (34 erroneous initial utterances vs 302 error free initial utterances) made a comparison unfeasible in the present case. Furthermore, the effect of error sources on users' error handling strategies and strategy rankings is already well understood [46, 97, 99].

Finally, the streamlined approach in terms of dialog turns and position of errors needs discussing. To draw comparable interactions and to account for the workload of the primary driving task, the number and position of errors was kept consistent. As costs have proven effective in determining the suitability of error handling strategies, we do not believe this approach restricts the generalizability of our results. On the contrary, it can be assumed that the importance of costs will only increase in multi-turn dialogs or scenarios with more than one error, where users have already invested more costs than at the outset of a dialog. Hence, we are confident that the concept of costs remains significant in these dialog settings. Nonetheless, future studies can build on our findings and examine the role of costs in multi-turn dialogs with various errors.

6 DESIGN IMPLICATIONS

The preceding paragraphs not only offer relevant insights for CUI designers but are also significant for CUI implementation.

Aiming at providing correct and context-sensitive *task-related* or *anticipatory* suggestions can be postulated as the indubitable best practice for repairing erroneous conversations: few turn-taking and word costs lead to quick and efficient error handling from the users' side. By implementing a *task-related* error handling strategy, missing information is queried in a direct and target-oriented manner. This strategy creates transparency for users. As they are led in how to repair a dialog, error spirals triggered from the users' side can be prevented. Furthermore, low word costs and the targeted retrieval of information support recognition from the CUI's side. *task-related* user repairs display an overall low number of turns and words, resulting in a low repair ratio. The repair ratio is an insightful means to quantify word-based repair costs. It describes the relation of number of words needed to repair a dialog in proportion to the words used to formulate the initial user utterance. The *task-related* strategy exhibited the overall lowest repair ratio and the only one with a ratio of below 50%. This low number of repair words supports recognition and can disrupt error spirals from the CUI's side. While open questions as in the *task-related* strategy can hence be advised, the following finding needs mentioning: more turns due to closed questions led to an overall lower number of words per turn. While strategies with closed questions were less popular among study participants, a stepwise inquiry of information as well as a low number of words per turn can support recognition. In these cases, fewer words need to be recognized and processed per dialog turn. This can be relevant for CUIs with unreliable recognition and/or security-relevant use cases, where correct recognition is especially crucial. In these cases, *anticipatory* error handling can be the apt design choice. Figure 9 visualizes concrete design recommendations in case correct suggestions are possible.

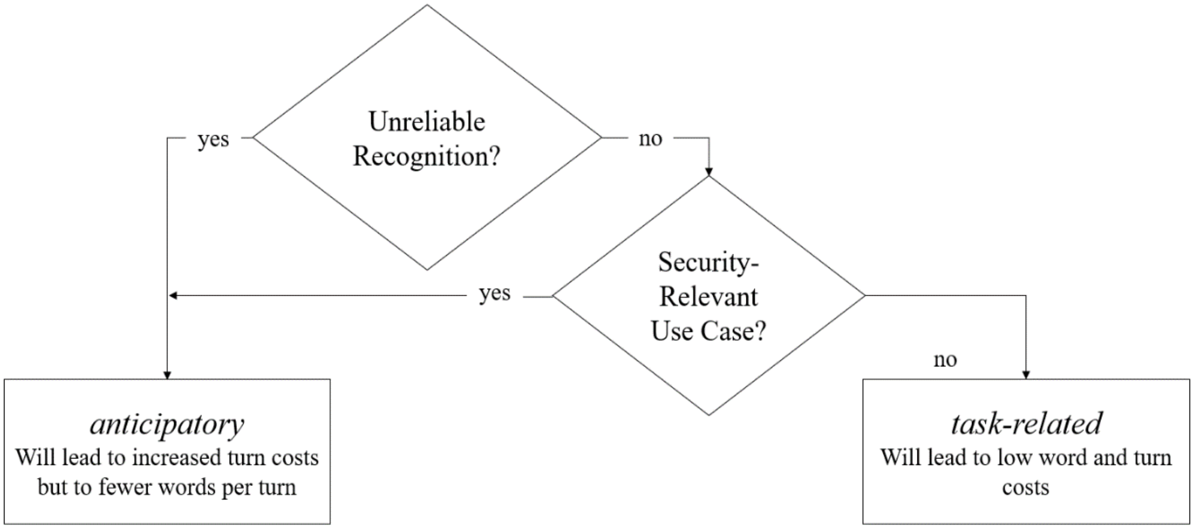

Unreliable Recognition?
yes
no
Security-Relevant Use Case?
yes
no
anticipatory
Will lead to increased turn costs but to fewer words per turn
task-related
Will lead to low word and turn costs

Figure 9: Design Guidelines for Repairing Conversational Breakdowns in Case of Correct Suggestions

In case an incorrect suggestion was made by the CUI, study participants preferred the baseline error handling strategy. Due to comparably low repair costs, asking task-related questions was ranked closely behind the baseline condition though. Anecdotally, study participants opted to repair baseline dialogs by repeating their initial utterances quasi verbatim. In case the underlying cause for an error is NLU-based rather than ASR-based, the baseline error handling strategy will lead to error spirals. Applying baseline error handling is hence only advised in case an error is not linked to a deficient NLU.

Full dialog resets were found to be a direct result of making incorrect suggestions, which led to significantly increased turn-taking costs. The necessity for a full reset thereby depended on the applied error handling strategy. task-related questions led to significantly fewer full resets than the strategies anticipatory and query. baseline dialogs can be classified as full resets per se, as they essentially restart a conversation. When designing error handling dialogs, conversational designers need to keep the occurrence of full dialog resets and their potential consequences like increased user dissatisfaction in mind. Although the baseline condition was preferred by study participants in case no correct suggestion could be made by the CUI, the strategy was found to be significantly longer than HHI-based strategies. Despite the user preference for baseline error handling, designers could opt to repair errors in time- and security-critical use cases with shorter strategies, such as task-related or anticipatory. Figure 10 shows design recommendations for error handling dialogs if no correct suggestion can be made by the CUI.

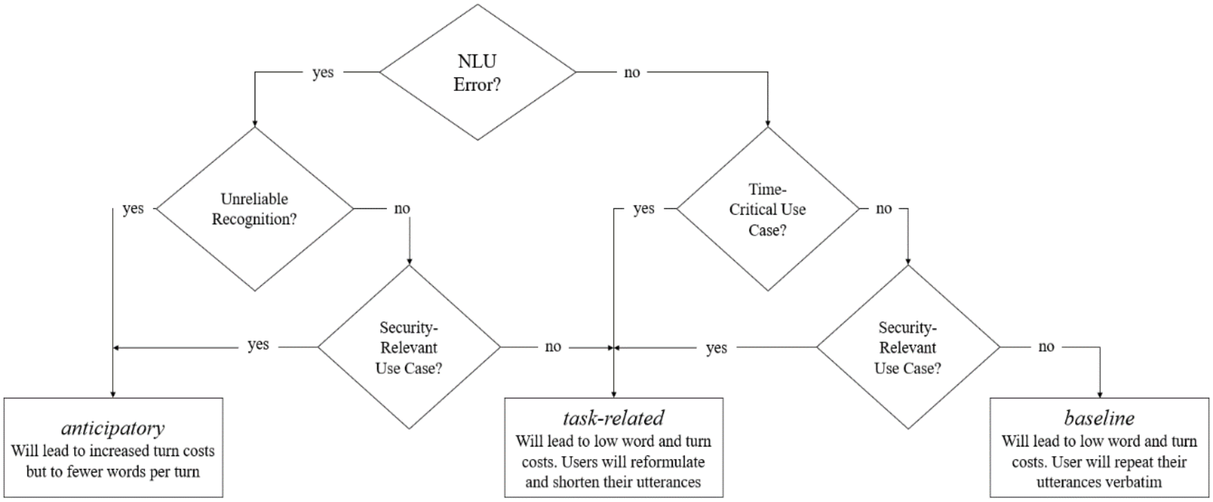

NLU Error?
yes
no
Unreliable Recognition?
yes
no
Time-Critical Use Case?
yes
no
Security-Relevant Use Case?
yes
no
Security-Relevant Use Case?
yes
no
anticipatory
Will lead to increased turn costs but to fewer words per turn
task-related
Will lead to low word and turn costs. Users will reformulate and shorten their utterances
baseline
Will lead to low word and turn costs. User will repeat their utterances verbatim

Figure 10: Design Guidelines for Repairing Conversational Breakdowns in Case of Incorrect Suggestions

7 CONCLUSION

In this study, three human-centred error handling strategies were developed and compared to a classic HCI error handling strategy for task-oriented dialogs in an in-car scenario. By means of a within-subjects driving simulator study in a high-fidelity driving simulator, the paper examined the applicability of the principle of Least Collaborative Effort to HCI. The principle of Least Collaborative Effort explains that conversational partners will work collaboratively towards the successful completion of a dialog while keeping so-called costs as low as possible. Costs can thereby be defined as the effort going along with leading successful and repairing failed interactions. While a whole catalogue of costs exists for HHI, the present research focused on word as well as turn-taking costs as indicators for error handling strategy preferences.

Quantitative as well as qualitative measures were collected from n=48 study participants to determine whether the principle of Least Collaborative Effort should be assigned a role in HCI. The user experience of different error handling strategies was thereby evaluated with the aid of the UEQ+ questionnaire items *response behavior* and *response quality* as well as through a qualitative strategy ranking. The UEQ+ only proved to be conclusive and significant with reservations. Costs calculations as well as qualitative strategy rankings on the other hand led to meaningful insights into user preferences for in-car error handling strategies as well as to derivations for CUI implementation.

Our findings show that human-centred error handling strategies lead to significantly lower repair costs than a classic HCI error handling strategy. Strategies with low word and turn-taking costs were thereby rated significantly better than strategies with higher costs. Strategies providing *task-related* questions or *anticipatory* suggestions were found to be least costly and hence most popular among study participants. The adequacy of suggestion, meaning the presentation of a context-relevant *task-related* or *anticipatory* proposal to solve an error, proved to be a decisive factor for a positive strategy ranking. In case no correct suggestion could be made by the CUI, study participants' strategy preferences shifted. Although *task-related* questions were still popular due to their low repair costs, *baseline* – meaning the classic HCI error handling strategy – was ranked best in this scenario. These findings complement prior research in the field of error handling in HCI, supporting that human-centred design approaches are necessary to facilitate ease of use and bridge current technological limitations. Our results show that preferences for error handling strategies are predictable by calculating strategies' concomitant repair costs in form of word and turn-taking costs. Costs can thereby be measured by adding up the number of turns and the number of words needed to repair an error. As such, costs are a straightforward and easily applicable as well as usable means to determine preferences for error handling strategies. Furthermore, considerations around word and turn-taking costs provide valuable insights for CUI implementation. Understanding the number of words and turns users need to repair errors can aid in supporting the strengths and weaknesses of a CUI's NLU.

The demonstrated correlation of costs and strategy preferences indicates that the principle of Least Collaborative Effort must be afforded a place in HCI. Furthermore, the obtained results introduce costs as a measurement tool for examining the suitability of (erroneous) dialogs and offer assistance for designing user-centric error handling strategies. In practice, this means that HCI practitioners and CUI designers can fall back on simple and straightforward cost calculations to design human-centred and low-cost error handling strategies.

4 CONCLUSION

4.1 Efficient Testing Methods for CUI Prompts

Testing methods for in-car CUIs span a broad spectrum between low- and high-fidelity studies. High-fidelity driving simulator studies fully immerse study participants in a driving task and are therefore, at first glance, the best choice when studying the user experience of in-car CUIs. However, driving simulator studies are highly time and cost consuming. Low-fidelity online crowdsourcing studies represent the opposite case: while a fraction of the cost and time is needed for their execution, they lack the immersive elements present in driving simulator studies. Still, the question arises if a high-fidelity environment is needed for the validation of single CUI elements, such as prompts. Taking a driving simulator study as the baseline, two low-fidelity crowdsourcing studies were conducted to examine how to efficiently validate in-car prompts. One of the studies presented CUI prompts in text form, while the other study introduced prompts as audio files. In a between-subjects design, 21 proactive prompts were evaluated by 75 study participants in both online conditions as well as by 58 study participants in the driving simulator condition. Prompts were rated more poorly in the audio condition than in the text and the simulator condition. Subsequent statistical analyses showed no differences in the evaluation of prompts in the text and the driving simulator condition. Evaluations in the audio condition on the other hand differed significantly from both text and simulator evaluations (see Figure 1: General Ratings Across Testing Conditions). As such, a significant impact of the testing condition on prompt evaluations was established.

Results are explained by the Elaboration Likelihood Model (ELM) [110]. The ELM identifies different so-called "routes" recipients of a message can take in this message's processing: a) the central or controlled, and b) the peripheral or automatic route of processing. The decision for a route is thereby made subconsciously and depends on a recipient's ability and motivation to process a message. If recipients can allocate "considerable cognitive resources" [110, p. 128] to processing a message, they will take the central route of processing. Else, they process a message based on its metadata rather than on the "true merits of the information presented" [110, p. 125]. This metadata may thereby contain the credibility of a source or the TTS voice presenting it. The obtained results suggest that study participants in the text and the simulator condition evaluate prompts by using the central route of processing, while they use the peripheral route in the audio condition (see Figure 3: Processing Routes Across Testing Conditions). These results prove that user experience testing for CUI prompts can be performed in low-fidelity study environments, such as online crowdsourcing studies. Thus, the research question of how to efficiently validate in-car prompts can be answered with this exact format.

4.2 Prompt Design Guidelines for Different Interaction Types

To close the gap of to date insufficient prompt design guidelines for CUIs, six online crowdsourcing studies were conducted. This study format was found to be an efficient and equally valid alternative to highly resource-intensive driving simulator studies. To gain insights into linguistic parameters with a potential influence on the evaluation of a prompt, a German contemporary grammar with 94 chapters was worked through [131]. In this step, 28 linguistic parameters – more specifically four syntactical, 13 grammatical, and 11 lexical parameters – were extracted (see Table 1: Overview over Syntactical, Grammatical, and Lexical Parameters). With this list, the paper's first research question, namely "What is the entirety of syntactical, grammatical, and lexical parameters with a potential impact on prompt design?" is answered.

The extracted parameters were subsequently cast into comparison prompts for functional, informational, as well as chit chat use cases. For each parameter, three comparison prompts were produced. Comparison prompts thereby varied in only one parameter at a time. As such, the final set of study prompts comprised 1044 prompts. Prompt pairs were presented to a total of 1206 study participants in an online crowdsourcing study in an A/B format. Paper 2, Table 2 "Overview over Results" gives an exhaustive overview over preferences for all examined linguistic parameters and answers the second research question "Which manifestation of syntactical, grammatical, and lexical parameters is preferred by participants for which prompt type?". The obtained results show that study participants do have linguistic preferences for the formulation of CUI prompts in an in-car setting. Moreover, these preferences partly vary depending on the type of conversation: of 28 examined parameters, nine parameters produced varying best practices across conversation types. As such, the type of conversation emerges as a context worth considering when designing in-car prompts.

By means of the compiled best practices, operationalized CUI-specific guidelines were formulated. Furthermore, clustering the obtained best practices led to the identification of user needs and overall design patterns for interactions with CUIs. With these results, the paper's two remaining research questions are answered. Clustering and interpretation of formulation preferences led to the discovery of three superordinate user needs for designing in-car prompts: 1) a suitable level of (in)formality, 2) a suitable level of complexity/simplicity, and 3) a suitable level of (im)mediacy. Regarding the level of (in)formality, study participants prefer balanced prompts with neither too informal nor too colloquial language. The level of complexity tilts towards simple language. Lastly, (im)mediacy describes study participants' observed need for straightforward formulations. Linguistic decorative attachments in form of e.g., a high adverb or modal particle density is deemed unnecessary by study participants and penalized in A/B studies. Based on these clusters, three main guidelines for designing in-car CUI prompts are formulated:

1. Prompts should be written in natural and rather informal language without being too colloquial.
2. Prompts should be written in plain and simple language to avoid complexity.
3. Prompts should be written results- and information-oriented, leaving out unnecessary elements.

Contrary to already existing prompt design guidelines, these guidelines are specifically tailored to in-car CUIs. Furthermore, they are empirically validated and can be adhered to in detail by assigning the obtained best practices to them (see Table 3: Best practices according to user needs). Parameters can thereby be mapped to guideline number one and the (in)formality dimension if they have both a written- and a spoken-language manifestation, such as an active vs a passive voice. Spoken-language manifestations are thereby less formal than their written-language counterparts. If parameters add to respectively reduce users' cognitive load, they can be counted to guideline number two and a suitable level of complexity/simplicity. This dimension is managed best on a syntactical level. Lastly, parameters are considered (im)mediate when they enable respectively halt direct formulations. While a high use of adverbs, adjectives, and modal particles purports natural language, it also inflates a prompt with unnecessary elements.

The obtained results are beneficial for CUI designers and practitioners in multiple ways. For one, the conversation type is identified as an important context in CUI prompt design. Secondly, large-scale linguistic-driven user studies reveal concrete, hands-on best practices for designing in-car CUI prompts on a syntactical, grammatical, and a lexical level. With aid of these best practices, three superordinate user needs are identified and lead to the formulation of three design guidelines. By mapping linguistic best practices onto these guidelines, they become an easily applicable and straightforward tool CUI designers can adhere too. As such, these findings close the gap of to date insufficient CUI design guidelines by providing a set of validated best practices for designing context-sensitive in-car prompts.

4.3 Prompt Design Guidelines for Different Interaction Domains

As established in the previous chapter, an integral part of a strong user experience is comprised of how CUIs talk to their users. Study participants show nuanced preferences for certain formulations on syntactical, grammatical, and lexical levels. These preferences thereby partly depend on the type of interaction. A related study dealing with syntactical considerations in prompt design [134] indicates that the topic of an interaction – the so-called interaction domain – needs to be considered as a further context for context-sensitive prompt design. Taking previously obtained results around interaction types as well as linguistic parameters with an impact on the evaluation of prompts into account, three domains were investigated: a) a security-relevant domain, b) a comfort-oriented domain, and c) a general functional domain. A within-subjects crowdsourcing study with 200 study participants was conducted to determine representative use cases for these domains. Subsequently, comparison prompts were designed across interaction types, linguistic parameters, and domains. Per domain, 200 study participants rated 54 CUI prompts in a between-subjects online crowdsourcing study (see Figure 2: Distribution and Structure of Crowdsourcing Studies).

Given the number of contexts considered, domain-sensitive linguistic preferences were hardly measurable. As such, domains are found to barely have an impact on linguistic preferences for CUI prompts. Only one linguistic parameter proves to be dependent on the type of conversation and its domain. Preferences for the position of sub-clauses in proactive interactions differs between a comfort-oriented and a functional domain. While prepositive sub-clauses are preferred in functional dialogs, postpositively put sub-clauses are study participants' choice for comfort-oriented interactions. However, the found effect size is rather small. Nonetheless, the result stresses the importance of proper information packaging in CUIs [26]. To present information in a manner that is appropriate for CUI users, the position of main and sub-clauses needs to be considered carefully. Sub-clauses are initiated by conjunctions which indicate what, when, why or how something is happening. As such, prepositive sub-clauses directly indicate the reason for a proactive interruption. Study participants' preference for this position in functional interactions shows that their need for information is higher in these cases than it is in comfort-related dialog. Still, the research question of which effect domains have on the preference for linguistic parameters across dialog types is best answered by "barely any". It can be concluded that domains are not the most important context to consider when designing CUI prompts. Rather, linguistic preferences for prompts are dependent on the type of interaction as shown in the previous chapter.

4.4 Prompt Design Guidelines for Proactive Interactions

With the increasing development from reactive to proactive CUIs, the systems make a leap towards reciprocal conversational interaction patterns [95]. While a multitude of studies has shown high approval ratings and user needs for proactive assistants [72, 141], research also speaks of a proactivity dilemma [148]. Proactive CUIs "need to strike the right balance between being helpful and being intrusive" [114, p. 2]. For instance, users may not be disturbed while carrying out primary tasks. This is especially important in the vehicle, where users are preoccupied with the potentially safety-critical primary task of driving. Guidelines for designing successful proactive interactions are hence particularly important in an in-car setting. Previous research around in-car proactivity has focused primarily on the right timing for proactive interactions. How to interrupt users on the other hand has received less attention. Considering this thesis' preceding findings regarding the importance of context-sensitive prompt design, proactivity emerges as an intriguing further interaction context. Hence, a study was designed to investigate clear linguistic-driven design guidelines for proactive in-car interactions, by answering the following research question: Are there best practices for the formulation of proactive prompts?

Building on related work, two linguistic dimensions for successful proactive interactions are identified: a) language complexity, and b) suggestive rather than imposing language [114, 148]. The previous papers produced a number of syntactical, grammatical, as well as lexical parameters with an impact on the evaluation of prompts. For the present study, these parameters were clustered onto the afore-mentioned linguistic dimensions (see Table 1: Overview and Clustering of Study Parameters). To test the user experience of single CUI prompts, the previous studies worked with online crowdsourcing studies in an A/B format. As the present study was conducted in a driving simulator, this testing method was not applicable. Hence, a Likert scale with the items *naturalness*, *simplicity*, *intelligence*, and *positivity* was developed based on already existing frameworks [67] and validated for the purpose of the study (see Figure 1: Likert Scale used in the Driving Simulator Study). Reliability tests were conducted with the aid of Cronbach's alpha and splithalf reliability tests, proving the proposed scale to be reliable. As such, it was adopted as the study scale for the driving simulator study. With this result, CUI designers and practitioners are provided with a validated measurement tool for examining the user experience of single CUI prompts.

Throughout the driving simulator study, two-thirds of the presented proactive suggestions were accepted by study participants. Immediate driving-related use cases were thereby rated better than comfort-related suggestions (see Table 4: Approval Rate of Proactive Use Cases). Demographic factors were found to only play a minor role in prompt evaluations. Although the present study was concerned with *how* to proactively interrupt drivers, these results fit well with already existing research regarding *when* to interrupt them. The obtained results identify language complexity as a decisive factor for prompt preferences in proactive in-car conversations. Language complexity can thereby be catered to by adhering to certain syntactical structures, such as paratactical sentences. Suggestive language on the other hand proves to be less formalizable for proactive prompts (see Table 6: Overview over Formulation Preferences for Proactive Prompts). As such, results regarding linguistic preferences differ between proactive interactions and previously examined types of conversations. While linguistic parameters prove to be important for functional, informational, and chit chat dialog, they are less pronounced in proactive interactions. Still, results underline the need to extend the existing design framework for proactive prompts by linguistic considerations.

4.5 Flow Design Guidelines for Conversational Breakdowns

While the previous chapters described best practices for designing CUI prompts for different conversational contexts, this chapter approaches guidelines for designing CUI flows. Decreasing word error rates and improved intent recognition have led to a leap in CUI performance. However, errors in conversations with CUIs are still commonplace [99] and oftentimes speaker dependent [54]. As such, designing suitable error handling strategies for CUIs represents an essential part of CUI's user experience. Erroneous conversations are thereby not exclusive to conversations in HCI but a normalcy in HHI too. Human conversational partners are therefore specialized in subconsciously repairing errors [16] with dedicated error handling strategies [31]. One of these strategies is the "Principle of Least Collaborative Effort" [31]. This principle expects interlocutors to repair errors context-sensitively and with a low amount of so-called costs. Costs thereby describe the effort needed to repair an error and steer a conversation towards successful completion. While they can span a broad range of factors, costs can be measured straightforwardly by counting the number of turns and the number of words needed to correct a faulty conversation. To gain insights into user preferences for error handling strategies and derive concrete guidelines for designing CUI flows, a driving simulator study was developed. In this study, 48 study participants experienced seven erroneous interactions which were repaired with different error handling strategies from the CUI’s side. As a baseline, these error handling strategies were compared to a classical HCI error handling strategy, namely *“I didn’t catch that. Can you please repeat that?”* This baseline strategy was contrasted against three human-centred error handling strategies: a) asking task-related questions, b) anticipatorily handling errors, and c) querying user input. For strategies a) to c), two scenarios were developed. In the first scenario, study participants were provided with a correct suggestion to repair the occurred error. In the second scenario, the provided suggestion did not match the error context and suggested a “false” solution to handle the error (see Table 1: Overview over Interactions). By means of quantitative as well as qualitative evaluations, the following two research questions were answered:

1. Do different error handling strategies vary regarding their repair costs?
2. Do the costs associated with repairing errors influence the preference for repair strategies?

The driving simulator study finds error handling strategies to differ significantly regarding their concomitant repair costs. HHI-centred error handling strategies are thereby significantly less costly than the compared HCI baseline strategy. More specifically, strategies a) asking task-related questions and b) anticipatorily handling errors lead to the most cost-efficient repairs (see Figure 6: Costs per Strategy). With this knowledge, CUI designers can design error handling flows in a cost-efficient manner. Moreover, the study finds a significant relationship between costs and preferences for error handling strategies. The lower the repair costs, the better a strategy is ranked by study participants in a qualitative ranking. Repair costs are thereby not only dependent on strategies per se, but also on the correctness of the proposed repair. For strategies with correct suggestions, human-centred error handling strategies are preferred over the HCI baseline strategy (see Figure 4: Turns, Words, and Total Costs for Error Handling Strategies with Correct Suggestions). In case a strategy includes an incorrect suggestion, repair costs for human-centred error handling strategies rise. Due to its nature, the baseline strategy is not affected by correct or incorrect suggestions as it merely asks study participants to repeat themselves. As such, the baseline condition is less costly than human-centred error handling strategies providing incorrect suggestions. In these cases, study participants’ strategy preferences shift to the less costly baseline condition. With this finding, research question two is answered, as findings demonstrate a significant relationship between preferences for error handling strategies and their concomitant repair costs (see Figure 5: Turns, Words, and Total Costs for Error Handling Strategies with Incorrect Suggestions). Study participants’ age, gender as well as their attitude towards CUIs do thereby not influence strategy preferences. Taken together, the obtained results underline the importance of context-sensitive CUI design once more. Further parameters prove important for high strategy rankings by study participants. Strategies asking open questions are repaired with significantly fewer words and turns than closed questions. As such, errors are repaired quicker when questions are posed as open questions. Moreover, clear confessions of an error lead to poorer strategy rankings. Although error confessions do not influence repair costs, strategies confessing errors are consistently ranked lower than strategies without such confessions.

FINAL SUMMARY

This thesis distils the importance of context-sensitive considerations when designing CUI prompts and flows in an in-car environment. These considerations can be cast into empirically obtained, concrete and directly applicable best practices and design guidelines, which are validated specifically for the CUI context. The following overview provides the answer to the research question of how to context-sensitively design CUI prompts and flows in a nutshell:

Testing Methods
Crowdsourcing
Driving Simulator

Make use of efficient online crowdsourcing studies to test the user experience of single CUI prompts.

Linguistics
Syntax
Grammar
Lexis

Consider linguistic best practices on syntactical, grammatical, and lexical levels. Prompts need to display a suitable level of (in)formality, complexity/simplicity, and (im)mediacy. They are to be written in natural and rather informal language without being too colloquial, be plain and simple, and results- and information-oriented.

Type of Interaction
Functional
Informational
Chit Chat

Consider the type of conversation when designing prompts. Distinguish between functional, informational, and chit chat dialog in terms of prompt length, modal particles, comparisons, position of sub-clauses, ellipses, and participles.

Domain of Interaction
Security-relevant
Comfort-oriented
Functional

Distinguish between functional and comfort-oriented dialog when designing proactive prompts. Use prepositive sub-clauses in functional, and postpositive sub-clauses in comfort-oriented proactive interactions. Domain-sensitivity is negligible in functional and informational dialog.

Initiation of Interaction
Proactive
Reactive

Design prompts with a low degree of linguistic complexity and a suggestive rather than imposing speaking style by adhering to concrete syntactical and lexical best practices. Consider *when* to proactively interrupt drivers before considering *how* to interrupt them.

Success of Interaction
Successful
Erroneous

Use different error handling strategies depending on the context. Calculate repair costs for error handling strategies by summing up the number of turns and the number of words needed to repair an error. Low repair costs correlate with high error handling strategy rankings.

FURTHER PUBLICATIONS

Further published papers and talks with reference to the dissertation:

1. PAPERS

- Meck, A.-M., Sardone, M., and Cullmann, J. 2023. Will the Assistant Become the Driver, and the Driver Become the Assistant? *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*, 1–5. https://doi.org/10.1145/3571884.3603753.
- Meck, A.-M. 2023. How May I Interrupt? Linguistic Design Guidelines for Proactive In-Car Voice Assistants. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, C. Draxler Ed. TUDpress, Dresden, 24-31. https://www.essv.de/paper.php?id=1169.
- Meck, A.-M., Edwards, J., Garaialde, D., Bartl, M., Doyle, P., and Clark, L. 2022. Design for the User You Want, Not the User You Have? *CUI@CHI: Ethics of Conversational User Interfaces: Virtual Workshop at the ACM CHI 2022 Conference*, 8 pages. https://www.conversationaluserinterfaces.org/workshops/CHI2022/pdfs/meck_Bias_Workshop_Paper.pdf.

2. TALKS

- 2nd Lunch Talk on Human-Centred Design at the Media Informatics and Human-Computer Interaction Groups, LMU Munich. Title: *"How Voice Assistants Should Talk"*.
- Doc/Post-Doc Colloquium at the Institute of Phonetics and Speech Processing, LMU Munich. Title: *"Failing with grace: Exploring Human Error Handling Strategies in Conversations with In-Car Voice Assistants"*.

REFERENCES

[1] Alvarez, I., Alnizami, H., Dunbar, J., Johnson, A., Jackson, F., and Gilbert, J. 2011. Designing Driver-Centric Natural Voice User Interfaces. *Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 156–159.

[2] Andolina, S., Orso, V., Schneider, H., Klouche, K., Ruotsalo, T., Gamberini, L., and Jacucci, G. 2018. Investigating Proactive Search Support in Conversations. *DIS '18: Proceedings of the 2018 Designing Interactive Systems Conference*, 1295–1307. https://doi.org/10.1145/3196709.3196734.

[3] Aneja, D., McDuff, D., and Czerwinski, M. 2020. Conversational Error Analysis in Human-Agent Interaction. *IVA ’20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA ’20)*, 8 pages. https://doi.org/10.1145/3383652.3423901.

[4] Ashktorab, Z., Jain, M., Liao, Q. V., and Weisz, J. D. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, 12 pages. https://doi.org/10.1145/3290605.3300484.

[5] Balters, S., Mauriello, M. L., Park, S. Y., Landay, J., and Paredes, P. O. 2020. Calm Commute: Guided Slow Breathing for Daily Stress Management in Drivers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4, 1, Article 38*, 19 pages. https://doi.org/10.1145/3380998.

[6] Barón, A. and Green, P. 2006. Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review. University of Michigan, Transportation Research Institute Ann Arbor, MI. https://api.semanticscholar.org/CorpusID:60137702.

[7] Bédard, M. B., Parkkari, M., Weaver, B., Riendeau, J., and Dahlquist, M. 2010. Assessment of driving performance using a simulator protocol: Validity and reproducibility. *The American Journal of Occupational Therapy 64*, 336–340. https://doi.org/10.5014/ajot.64.2.336.

[8] Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., and Yip, J. 2019. Communication Breakdowns Between Families and Alexa. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 13 pages. https://doi.org/10.1145/3290605.3300473.

[9] Biermann, M., Schweiger, E., and Jentsch, M. 2019. Talking to Stupid?!? Improving Voice User Interfaces. In *Mensch und Computer 2019 - Usability Professionals*, H. Fischer and S. Hess, Eds. Gesellschaft für Informatik e.V. Und German UPA e.V., Bonn, 53–61. https://doi.org/10.18420/muc2019-up-0253.

[10] Blair, J. and Abdullah, S. 2020. "It Didn’t Sound Good with My Cochlear Implants": Understanding the Challenges of Using Smart Assistants for Deaf and Hard of Hearing Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 4, Issue 4*, 27 pages. https://doi.org/10.1145/3432194.

[11] Bohus, D. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*. Doctoral dissertation, Carnegie Mellon University, School of Computer Science, Defense Technical Information Center. https://apps.dtic.mil/sti/pdfs/ADA476845.pdf.

[12] Bohus, D. and Rudnicky Alexander I. 2005. Sorry I didn't Catch That: An Investigation of Non-understanding Errors and Recovery Strategies. *Proceedings of the 6th SIGdial*

Workshop on Discourse and Dialogue, 16 pages. https://doi.org/10.1007/978-1-4020-6821-8_6.

[13] Branham, S. M. and Mukkath Roy, A. R. 2019. Reading Between the Guidelines: How Commercial Voice Assistant Guidelines Hinder Accessibility for Blind Users. *ASSETS'19: The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 446–458. https://doi.org/10.1145/3308561.3353797.

[14] Braun, M., Mainz, A., Chadowitz, R., Pfleging, B., and Alt, F. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. *CHI'19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290605.3300270.

[15] Brennan, S. E. 1998. The Grounding Problem in Conversations With and Through Computers. In *Social and cognitive psychological approaches to interpersonal communication*, S. R. Fussell and R. J. Kreuz, Eds. Lawrence Erlbaum, Hillsdale, NJ, 201-225.

[16] Brinton, B., Fujiki, M., Loeb, D. F., and Winkler, E. 1986. Development of Conversational Repair Strategies in Response to Requests for Clarification. *Journal of Speech, Language, and Hearing Research*, 75–81. https://doi.org/10.1044/jshr.2901.75.

[17] Brooke, J. 1995. SUS: A quick and dirty usability scale. In *Usability Evaluation In Industry*, P. W. Jordan, B. Thomas, I. L. McClelland and B. Weerdmeester, Eds. CRC Press, Boca Raton, FL, 8 pages.

[18] Brown, P. and Levinson, S. C. 1987. *Politeness: Some Universals in Language Use*. Cambridge Universal Press, Cambridge, MA.

[19] Brysbaert, M. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language Volume 109*, 94 pages. https://doi.org/10.1016/j.jml.2019.104047.

[20] Cabral, J., Cowan, B. R., Zibrek, K., and McDonnell, R. 2017. The Influence of Synthetic Voice on the Evaluation of a Virtual Character. *Interspeech 2017*, 229–233. https://doi.org/10.21437/Interspeech.2017-325.

[21] Cahn, J. E. and Brennan, S. E. 1999. A Psychological Model of Grounding and Repair in Dialog. *AAAI Technical Report FS-99-03*, 9 pages. http://www.psychology.sunysb.edu/sbrennan-/papers/cahnbren.pdf.

[22] Callender, J. C. and Osburn, H. G. 1979. An empirical comparison of Coefficient Alpha, Guttman’s Lambda – 2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement 16(2)*, 89–99. https://doi.org/10.1111/j.1745-3984.1979.tb00090.x.

[23] Cambre, J. and Kulkarni, C. 2020. Methods and Tools for Prototyping Voice Interfaces. *2nd Conference on Conversational User Interfaces (CUI'20)*, 1–4. https://doi.org/10.1145/3405755.3406148.

[24] Capgemini Research Institute. 2019. Voice on the go. How can auto manufacturers provide a superior in-car voice experience. Capgemini Research Institute. Retrieved May 13, 2021 from https://www.capgemini.com/de-de/wp-content/uploads/sites/5/2019/11/Report-%E2%80%93-Voice-on-the-Go.pdf.

[25] Cha, N., Kim, A., Park, C. Y., Kang, S., Park, M., Lee, J.-G., Lee, S., and Lee, U. 2020. “Hello There! Is Now a Good Time to Talk?”: Opportune Moments for Proactive Interactions with Smart Speakers. *Proceedings of the ACM on Interactive, Mobile,*

Wearable and Ubiquitous Technologies, Volume 4, Issue 3, 74, 1–28. https://doi.org/10.1145/3411810.

[26] Chafe, W. and Danielewicz, J. 1987. Properties of spoken and written language. In *Comprehending oral and written language*, R. Horowitz and S. J. Samuels, Eds. Academic Press, Cambridge, 83–113.

[27] Cheng, Y., Yen, K., Chen, Y., Chen, S., and Hiniker, A. 2018. Why Doesn't It Work? Voice-Driven Interfaces and Young Children's Communication Repair Strategies. *IDC '18: Proceedings of the 17th ACM Conference on Interaction Design and Children*, 337–348. https://doi.org/10.1145/3202185.3202749.

[28] Chérif, E. and Lemoine, J.-F. 2019. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Recherche et Applications en Marketing 34(1)*, 28–47. doi.org/10.1177/2051570719829432.

[29] Choi, W., Park, S., Kim, D., Lim, Y.-K., and Lee, U. 2019. Multi-Stage Receptivity Model for Mobile Just-In-Time Health Intervention. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 2, Article 39*, 26 pages. https://doi.org/10.1145/3328910.

[30] Christensen, R. H. B. 2019. *ordinal---Regression Models for Ordinal Data*. https://cran.r-project.org/web/packages/ordinal/ordinal.pdf.

[31] Clark, H. H. and Brennan, S. E. 1991. Grounding in Communication. In *Perspectives on socially shared cognition.*, L. B. Resnick, J. M. Levine and S. D. Teasley, Eds. APA Books, Washington, 127–149.

[32] Clark, L. 2018. Social boundaries of appropriate speech in HCI: A politeness perspective. *Proceedings of the 32nd International BCS Human Computer Interaction Conference*, 1–5. https://doi.org/10.14236/ewic/HCI2018.76.

[33] Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., and Cowan, B. R. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 12 pages. https://doi.org/10.1145/3290605.3300705.

[34] Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Early, D., and Bandeira, N. 2017. "What Can I Help You With?": Infrequent Users' Experiences of Intelligent Personal Assistants. *MobileHCI '17: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–12. https://doi.org/10.1145/3098279.3098539.

[35] Cuadra, A., Li, S., Lee, H., Cho, J., and Ju, W. 2021. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proceedings of the ACM on Human-Computer Interaction Volume 5 Issue CSCW1*, 1–24. https://doi.org/10.1145/3449101.

[36] Defined.ai. 2023, Seattle, WA. https://www.defined.ai/.

[37] Demaree, H., Shenal, B., Everhart, D. E., and Robinson, J. 2004. Primacy and Recency Effects Found Using Affective Word Lists. *Cognitive and Behavioral Neurology 17(2)*, 102–108. https://doi.org/10.1097/01.wnn.0000117861.44205.31.

[38] Demberg, V., Sayeed, A., Mahr, A., and Müller, C. 2013. Measuring linguistically-induced cognitive load during driving using the ConTRe task. *AutomotiveUI '13: Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 176–183. https://doi.org/10.1145/2516540.2516546.

[39] DIN EN ISO 9241-110. 2020. Ergonomics of human-system interaction - Part 110: Interaction principles (ISO 9241-110:2020). https://doi.org/10.31030/3147467.

[40] Ekman, P. and Friesen, W. 1969. The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica, vol. 1, no. 1*, 49–98. https://doi.org/10.1515/semi.1969.1.1.49.

[41] Engelhardt, S., Hansson, E., and Leite, I. 2017. Better Faulty than Sorry: Investigating Social Recovery Strategies to Minimize the Impact of Failure in Human-Robot Interaction. *Proceedings of the first Workshop on Conversational Interruptions in Human-Agent Interactions co-located with 17th International Conference on International Conference on Intelligent Virtual Agents (IVA 2017)*, 19–27. http://ceur-ws.org/Vol-1943/WCIHAI-17-03.pdf.

[42] Engonopulos, N., Sayeed, A., and Demberg, V. 2013. Language and cognitive load in a dual task environment. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35(35)*, 2148–2153. https://escholarship.org/content/qt8p586904/qt8p586904.pdf.

[43] Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. 2021. *G * Power*. Computer Software, Heinrich Heine Universität Düsseldorf.

[44] Féry, C. and Ishihara, S. 2016. Introduction. In *The Oxford Handbook of Information Structure*, C. Féry and S. Ishihara, Eds. Oxford University Press, Oxford, 1–15.

[45] Furqan, A., Myers, C., and Zhu, J. 2017. Learnability through Adaptive Discovery Tools in Voice User Interfaces. *CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1617–1623. https://doi.org/10.1145/3027063.3053166.

[46] Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., and Tscheligi, M. 2015. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in Psychology, 6*, 12 pages. 10.3389/fpsyg.2015.00931.

[47] Gollasch, D. and Weber, G. 2021. Age-Related Differences in Preferences for Using Voice Assistants. *MuC '21: Proceedings of Mensch und Computer 2021*, 156–167. https://doi.org/10.1145/3473856.3473889.

[48] Greenwald, A. 1976. Within-Subject Designs: To Use Or Not To Use? *Psychological Bulletin Vol. 83, No. 2*, 314–320. https://doi.org/10.1037/0033-2909.83.2.314.

[49] Grice, P. 1975. Logic and Conversation. In *Syntax & Semantics*, P. Cole and J. L. Morgan, Eds. Academic Press, 41–58.

[50] Gutzmann, D. and Turgay, K. 2016. Zur Stellung von Modalpartikeln in der gesprochenen Sprache. *Deutsche Sprache, 44.2*, 97–122. https://doi.org/10.37307/j.1868-775X.2016.02.02.

[51] Haas, G., Rietzler, M., Jones, M., and Rukzio, E. 2022. Keep it Short: A Comparison of Voice Assistants' Response Behavior. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, 12 pages. https://doi.org/10.1145/3491102.3517684.

[52] Hagen, A. 2014. Zum Beispiel im Deutschen: Wenn Präposition und Artikel verschmelzen. Ein Korpusuntersuchung zu auf + definitem Artikel. *Sprachreport. Informationen und Meinungen zur deutschen Sprache*, 14–21. https://pub.ids-mannheim.de//laufend/sprachreport/pdf/sr14-3b.pdf.

[53] Harlann, I. 2017. *Voice interfaces are here to stay*. Retrieved November 2, 2022 from https://becominghuman.ai/voice-interfaces-are-here-to-stay-f2d3d206a6c4.

[54] Harrington, C. N., Garg, R., Woodward, A., and Williams, D. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health

Information Seeking. *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3491102.3501995.

[55] Hinterleitner, F. 2017. Influencing Factors on Perceptual Quality. *Quality of Synthetic Speech. T-Labs Series in Telecommunication Services*, 69–100. https://doi.org/10.1007/978-981-10-3734-4_5.

[56] Hofmann, H., Hermanutz, M., Tobisch, V., and Ehrlich, U. 2016. Evaluation of In-Car SDS Notification Concepts for Incoming Proactive Events. In *Situated Dialog in Speech-Based Human-Computer Interaction. Signals and Communication Technology*, A. Rudnicky, A. Raux, I. Lane and T. Misu, Eds. Springer, Cham, 102–112.

[57] Hone, K. and Graham, R. 2001. Subjective Assessment of Speech-System Interface Usability. *7th European Conference on Speech Communication and Technology*, 4 pages. https://doi.org/10.21437/Eurospeech.2001-491.

[58] Hwangbo, S. W., Classen, S., Mason, J., Yang, W., McKinney, B., Kwan, J., and Sisiopiku Virginia. 2022. Predictors of Simulator Sickness Provocation in a Driving Simulator Operating in Autonomous Mode. *Safety 2022, 8(4), 73*, 12 pages. https://doi.org/10.3390/safety8040073.

[59] Iqbal, S. T. and Bailey, B. T. 2005. Investigating the Effectiveness of Mental Workload as a Predictor of Opportune Moments for Interruption. *CHI EA '05: CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1489–1492. https://doi.org/10.1145/1056808.1056948.

[60] Jain, M., Kumar, P., Kota, R., and Patel, S. N. 2018. Evaluating and Informing the Design of Chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference (DIS ’18)*, 895–906. https://doi.org/10.1145/3196709.3196735.

[61] Kaiser, C. and Schallner, R. 2022. The impact of emotional voice assistants on consumers’ shopping attitude and behavior. *Wirtschaftsinformatik 2022*, 5 pages. https://aisel.aisnet.org/wi2022/workshops/workshops/10.

[62] Kim, A., Choi, W., Park, J., Kim, K., and Lee, U. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-vehicle Proactive Auditory-verbal Tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 2, Issue 4*, 28 pages. https://doi.org/10.1145/3287053.

[63] Kim, A., Park, J.-M., and Lee, U. 2020. Interruptibility for In-vehicle Multitasking: Influence of Voice Task Demands and Adaptive Behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 4, Issue 1*, 22 pages. https://doi.org/10.1145/3381009.

[64] Kim, J., Jeong, M., and Lee, S. C. 2019. "Why did this voice agent not understand me?": error recovery strategy for in-vehicle voice user interface. *AutomotiveUI '19: Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 146–150. https://doi.org/10.1145/3349263.3351513.

[65] Kintsch, W. and van Dijk, T. A. 1978. Toward a model of text comprehension and production. *Psychological Review, 85(5)*, 363–394. https://doi.org/10.1037/0033-295X.85.5.363.

[66] Kisler, T., Reichel, U. D., and Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language, Volume 45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005.

[67] Klein, A., Hinderks, A., Schrepp, M., and Tomaschewski, J. 2020. Construction of UEQ+ Scales for Voice Quality: Measuring user experience quality of voice interaction. *MuC '20: Proceedings of Mensch und Computer 2020*, 1–5. https://doi.org/10.1145/3404983.3410003.

[68] Klemmer, S. R., Sinha, A. K., Chen, J., Landay, J. A., Aboobaker, N., and Wang, A. 2000. Suede: A Wizard of Oz prototyping tool for speech user interfaces. *Proceedings of the 13th annual ACM symposium on User interface software and technology*, 1–10. https://doi.org/10.1145/354401.354406.

[69] Koch, K., Mishra, V., Liu, S., Berger, T., Fleisch, E., Kotz, D., and Wortmann, F. 2021. When Do Drivers Interact with In-Vehicle Well-being Interventions? An Exploratory Analysis of a Longitudinal Study on Public Roads. *Proc ACM Interact Mob Wearable Ubiquitous Technol. 5(1)*, 48 pages. https://doi.org/10.1145/3448116.

[70] Koch, K., Tiefenbeck, V., Liu, S., Berger, T., Fleisch, E., and Wortmann, F. 2021. Taking Mental Health &Well-Being to the Streets: An Exploratory Evaluation of In-Vehicle Interventions in the Wild. *CHI Conference on Human Factors in Computing Systems (CHI'21)*, 15 pages. https://doi.org/10.1145/3411764.3446865.

[71] Kontogiorgos, D., van Waveren, S., Wallberg, O., Pereira, A., Leite, I., and Gustafson, J. 2020. Embodiment Effects in Interactions with Failing Robots. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376372.

[72] Kraus, M., Wagner, N., Callejas, Z., and Minker, W. 2021. The Role of Trust in Proactive Conversational Assistants. *IEEE Access, vol. 9*, 112821–112836. https://doi.org/10.1109/ACCESS.2021.3103893.

[73] Kühne, K., Fischer, M. H., and Zhou, Y. 2020. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Front. Neurorobot.*, 15 pages. https://doi.org/10.3389/fnbot.2020.593732.

[74] Large, D. R., Burnett, G., and Clark, L. 2019. Lessons from Oz: design guidelines for automotive conversational user interfaces. *AutomotiveUI '19: Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 335–340. https://doi.org/10.1145/3349263.3351314.

[75] Large, D. R., Clark, L., Quandt, A., Burnett, G., and Skrypchuk, L. 2017. Steering the conversation: a linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied ergonomics* 63, 53–61. https://doi.org/10.1016/j.apergo.2017.04.003.

[76] Laugwitz, B., Held, T., and Schrepp, M. 2008. Construction and Evaluation of a User Experience Questionnaire. *HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society*, 63–76. https://doi.org/10.1007/978-3-540-89350-9_6.

[77] Lee, E.-J. 2010. The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Computers in Human Behavior, Volume 26, Issue 4*, 665–672. https://doi.org/10.1016/j.chb.2010.01.003.

[78] Lee, J.-G. and Lee, K. M. 2022. Polite speech strategies and their impact on drivers' trust in autonomous vehicles. *Computers in Human Behavior, Volume 127, 107015*. https://doi.org/10.1016/j.chb.2021.107015.

[79] Lee, M. K., Kiesler, S., Forlizz, J., Srinivasa, S., and Rybski, P. 2010. Gracefully mitigating breakdowns in robotic services. *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 8 pages. https://doi.org/10.1109/HRI.2010.5453195.

[80] Lee, U., Han, K., Cho, H., Chung, K.-M., Hong, H., Lee, S.-J., Noh, Y., Park, S., and Carroll, J. M. 2019. Intelligent positive computing with mobile, wearable, and IoT devices: Literature review and research directions. *Ad Hoc Networks 83*, 8–24. https://doi.org/10.1016/j.adhoc.2018.08.021.

[81] Lenerz, J. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. TBL-Verlag Narr, Tübingen.

[82] Lenhard, W. and Lenhard, A. 2011. Berechnung des Lesbarkeitsindex LIX nach Björnson. *Psychometrica*. https://doi.org/10.13140/RG.2.1.1512.3447.

[83] Limerick, H., Coyle, D., and Moore, J. W. 2015. Empirical Evidence for a Diminished Sense of Agency in Speech Interfaces. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*, 3967–3970. https://doi.org/10.1145/2702123.2702379.

[84] Linke, A., Nussbaumer, M., and Portmann, P. R., Eds. 2004. *Studienbuch Linguistik*. Niemeyer, Tübingen.

[85] Linnemann, G. A. and Jucks, R. 2018. ‘Can I Trust the Spoken Dialogue System Because It Uses the Same Words as I Do?’—Influence of Lexically Aligned Spoken Dialogue Systems on Trustworthiness and User Satisfaction. *Interacting with Computers, Volume 30, Issue 3*, 173–186. https://doi.org/10.1093/iwc/iwy005.

[86] Luger, E. and Sellen, A. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288.

[87] Mahmood, A., Fung, J. W., Won, I., and Huang, C.-M. 2022. Owning Mistakes Sincerely: Strategies for Mitigating AI Errors. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, 1–11. https://doi.org/10.1145/3491102.3517565.

[88] Malone, S. and Brünken, R. 2021. Hazard Perception, Presence, and Simulation Sickness—A Comparison of Desktop and Head-Mounted Display for Driving Simulation. *Front. Psychol. 12:647723*, 15 pages. https://doi.org/10.3389/fpsyg.2021.647723.

[89] Mavrina, L., Szczuzka, J., Strathmann, C., Bohnenkamp, L. M., Krämer, N., and Kopp, S. 2022. “Alexa, You're Really Stupid”: A Longitudinal Field Study on Communication Breakdowns Between Family Members and a Voice Assistant. *Frontiers in Computer Science, Sec. Human-Media Interaction*, 16 pages. https://doi.org/10.3389/fcomp.2022.791704.

[90] McElroy, K. 2017. *Prototyping for Designers: Developing the Best Digital and Physical Products*. O'Reilly UK Ltd., Sebastopol, CA.

[91] McGinn, C. and Torre, I. 2019. Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots. *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 211–221. https://doi.org/10.1109/HRI.2019.8673305.

[92] McLean, G. and Osei-Frimpong, K. 2019. Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior Volume 99*, 28–37. https://doi.org/10.1016/j.chb.2019.05.009.

[93] Meck, A.-M., Draxler, C., and Vogt, T. 2022. A Question of Fidelity: Comparing Different User Testing Methods for Evaluating In-Car Prompts. *Proceedings of the 4th Conference on Conversational User Interfaces (CUI'22)*, 1–5. https://doi.org/10.1145/3543829.3544519.

[94] Meck, A.-M. and Precht, L. 2021. How to Design the Perfect Prompt: A Linguistic Approach to Prompt Design in Automotive Voice Assistants – An Exploratory Study. *Proceedings of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '21)*, 237–246. https://doi.org/10.1145/3409118.3475144.

[95] Miksik, O., Munasinghe, I., Asensio-Cubero, J., Reddy Bethi, S., Huang, S. T., Zylfo, S., Liu, X., Nica, T., Mitrcsak, A., Mezza, S., Beard, R., Shi, R., Ng, R. W., Mediano, P., Fountas, Z., Lee, S. H., Medvesek, J., Zhuang, H., Rogers, Y., and Swietojanski, P. 2020. Building Proactive Voice Assistants: When and How (not) to Interact. *ArXiv*, 17 pages. https://arxiv.org/pdf/2005.01322.pdf.

[96] Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., and Tscheligi, M. 2017. To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI*, 15 pages. https://doi.org/10.3389/frobt.2017.00021.

[97] Motta, I. and Quaresma, M. 2021. Users' Error Recovery Strategies in the Interaction with Voice Assistants (VAs). In *Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021). Lecture Notes in Networks and Systems*, N. L. Black, W. P. Neumann and I. Noy, Eds. Springer, Cham, 658–666.

[98] Murad, C., Munteanu, C., Clark, L., and Cowan, B. R. 2018. Design guidelines for hands-free speech interaction. *MobileHCI '18: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, 269–276. https://doi.org/10.1145/3236112.3236149.

[99] Myers, C., Furqan, A., Nebolsky, J., Caro, K., and Zhu, J. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3173574.3173580.

[100] Nallapaneni, A. 2021. *Identifying the Influence of Emotional Voice Style in Proactive Automobile Voice Interfaces*. Master's thesis, Eindhoven University of Technology (TU/e). https://pure.tue.nl/ws/portalfiles/portal/170468358/1297864_Nallapaneni.pdf.

[101] Nass, C. and Brave, S. 2007. *Wired for Speech*. The MIT Press, Cambridge, MA.

[102] Nielsen, J. and Molich, R. 1990. Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*, 249–256. https://doi.org/10.1145/97243.97281.

[103] Norman, D. 2013. *The Design of Everyday Things*. Basic Books, New York, NY.

[104] Nothdurft, F., Ultes, S., and Minker, W. 2014. Finding Appropriate Interaction Strategies for Proactive Dialogue Systems—An Open Quest. *Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication*, 73–80.

[105] Paredes, P. O., Zhou, Y., Al-Huda Hamdan, N., Balters, S., Murnane, E., Ju, W., and Landay, J. 2018. Just Breathe: In-Car Interventions for Guided Slow Breathing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2,* 23 pages. https://doi.org/10.1145/3191760.

[106] Park, D., Yoon, W. C., and Lee, U. 2020. Cognitive States Matter: Design Guidelines for Driving Situation Awareness in Smart Vehicles. *Sensors 2020, 20, 2978*, 26 pages. https://doi.org/10.3390/s20102978.

[107] Pauzie, A. 2008. A method to assess the driver mental workload: the Driving Activity Load Index (DALI). *IET Intelligent Transport Systems Vol 2*, 315–322. https://doi: 10.1049/iet-its:20080023.

[108] Pearl, C. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences*. O'Reilly Media, Sebastopol.

[109] Pejovic, V. and Musolesi, M. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. *UbiComp '14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 897–908. https://doi.org/10.1145/2632048.2632062.

[110] Petty, R. E. and Cacioppo, J. T. 1986. The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology 19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2.

[111] Porcheron, M., Fischer, J., and Sharples, S. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 207–219. https://doi.org/10.1145/2998181.2998298.

[112] Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. 2018. Voice Interfaces in Everyday Life. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3173574.3174214.

[113] Pradhan, A., Mehta, K., and Findlater, L. 2018. Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 13 pages. https://doi.org/10.1145/3173574.3174033.

[114] Reicherts, L., Zargham, N., Bonfert, M., Rogers, Y., and Malaka, R. 2021. May I Interrupt? Diverging Opinions on Proactive Smart Speakers. *CUI '21: 3rd Conference on Conversational User Interfaces*, 1-10. https://doi.org/10.1145/3469595.3469629.

[115] Rodero, E. 2019. Do Your Ads Talk Too Fast To Your Audio Audience? How Speech Rates of Audio Commercials Influence Cognitive and Physiological Outcomes. *Journal of Advertising Research 60 (3)*, 337–349. doi.org/10.2501/JAR-2019-038.

[116] RStudio Team. 2022. *RStudio: Integrated Development Environment for R*, RStudio.

[117] Rudd, J., Stern, K., and Isensee, S. 1996. Low vs. high-fidelity prototyping debate. *Interactions Volume 3 Issue 1*, 76–85. https://doi.org/10.1145/223500.223514.

[118] Scates, K. 2022. *Voice Assistants Become More Proactive and 8 Other Predictions for 2022*. Retrieved January 11, 2022 from https://www.soundhound.com/voice-ai-blog/voice-assistants-become-more-proactive-and-8-other-predictions-for-2022/.

[119] Schegloff, E. A., Jefferson, G., and Sacks, H. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language 53(2)*, 361–382. https://doi:10.2307/413107.

[120] Schmidt, M., Bhandare, O., Prabhune, A., Minker, W., and Werner, S. 2020. Classifying Cognitive Load for a Proactive In-Car Voice Assistant. *IEEE Sixth International Conference on Big Data Computing Service and Applications*, 9–16. https://doi.org/10.1109/BigDataService49289.2020.00010.

[121] Schmidt, M. and Braunger, P. 2018. A Survey on Different Means of Personalized Dialog Output for an Adaptive Personal Assistant. *UMAP '18: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, 75–81. https://doi.org/10.1145/3213586.3226198.

[122] Schmidt, M., Minker, W., and Werner, S. 2020. How Users React to Proactive Voice Assistant Behavior While Driving. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 485–490. https://aclanthology.org/2020.lrec-1.61.pdf.

[123] Schmidt, M., Minker, W., and Werner, S. 2020. User Acceptance of Proactive Voice Assistant Behavior. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, 18–25. https://www.essv.de/pdf/2020_18_25.pdf.

[124] Schmidt, M., Stier, D., Werner, S., and Minker, W. 2019. Exploration and assessment of proactive use cases for an in-car voice assistant. *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, 148–155. https://www.essv.de/paper.php?id=76.

[125] Schrepp, M., Hinderks, A., and Tomaschewski, J. 2017. Konstruktion einer Kurzversion des User Experience Questionnaire. In *Mensch und Computer*, M. Burghardt, R. Wimmer, C. Wolff and C. Womser-Hacker, Eds., 5 pages.

[126] Schwartz, E. H. 2021. *Cerence Co-Pilot Introduces Proactive Car Voice Assistant*. Retrieved March 12, 2023 from https://voicebot.ai/2021/12/22/cerence-co-pilot-introduces-proactive-car-voice-assistant/.

[127] Semmens, R., Martelaro, N., Kaveti, P., Stent, S., and Ju, W. 2019. Is Now A Good Time?: An Empirical Study of Vehicle-Driver Communication Timing. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 12 pages. https://doi.org/10.1145/3290605.3300867.

[128] Sgall, P. 2015. Zur Stellung der Thema-Rhema-Gliederung in der Sprachbeschreibung. In *Papers on functional sentence perspective*, F. Danes, Ed. De Gruyter Mouton, Berlin, 54–74.

[129] Shneiderman, B. and Plaisant, C. 2004. *Designing the user interface: Strategies for effective Human-Computer Interaction*. Addison Wesley, Boston, MA.

[130] Skantze, G. 2007. Error Handling in Spoken Dialogue Systems: Managing Uncertainty, Grounding and Miscommunication. Doctoral dissertation, KTH Computer Science and Communication, Department of Speech, Music and Hearing. ISCA Archive. https://www.isca-speech.org/archive_open/archive_papers/ehsd2003/ehsd_071.pdf.

[131] Sommerfeldt, K.-E., Starke, G., and Hackel, W. 2011. *Einführung in die Grammatik der deutschen Gegenwartssprache*. Max Niemeyer Verlag, Tübingen.

[132] Stier, D., Heid, U., Kittel, P., Schmidt, M., and Minker, W. 2020. The Influence of Syntax on the Perception of In-Vehicle Prompts and Driving Performance. In *Conversational Dialogue Systems for the Next Decade*, L. F. D'Haro, Z. Callejas and S. Nakamuara, Eds. Springer, Singapur, 349–362.

[133] Stier, D., Heid, U., and Minker, W. 2020. Adapting In-Vehicle Voice Output: A User- and Situation-Adaptive Approach. *AutomotiveUI '20: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 12–15. https://doi.org/10.1145/3409251.3411711.

[134] Stier, D. and Sigloch, E. 2019. Linguistic Design of In-Vehicle Prompts in Adaptive Dialog Systems: An Analysis of Potential Factors Involved in the Perception of Naturalness. *UMAP '19: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 191–195. https://doi.org/10.1145/3320435.3320469.

[135] Strayer, David L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., and Biondi, F. 2015. Assessing Cognitive Distraction in the Automobile. *The Journal of the*

Human Factors and Ergonomics Society, 1300–1324. https://doi.org/10.1177/0018720815575149.

[136] Streiner, D. L. 2003. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment, 80:1*, 99–103. https://doi.org/10.1207/S15327752JPA8001_18.

[137] Suhm, B., Myers, B., and Waibel, A. 2001. Multimodal Error Correction for Speech User Interfaces. *ACM Transactions on Computer-Human Interaction, Vol. 8, No. 1*, 60–98. https://doi.org/10.1145/371127.371166.

[138] Tiong, E., Seow, O., Camburn, B., Teo, K., Silva, A., Wood, K. L., Jensen, D. D., and Yang, M. C. 2019. The Economies and Dimensionality of Design Prototyping: Value, Time, Cost, and Fidelity. *Journal of Mechanical Design 141(3)*, 18 pages. doi.org/10.1115/1.4042337.

[139] Vinciarelli, A., Pantic, M., and Bourlard, H. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing, Volume 27, Issue 12*, 1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007.

[140] Vlahos, J. 2019. *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt, Boston, MA.

[141] Völkel, S. T., Buschek, D., Eiband, M., Cowan, B. R., and Hussmann, H. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI'21)*, 1–15. https://doi.org/10.1145/3411764.3445536.

[142] Watzlawick, P., Beavin, J. H., and Jackson, D. D. 2017. *Menschliche Kommunikation. Formen, Störungen, Paradoxien*. hogrefe, Bern.

[143] Winter, J. C. de, Groot, S. de, Mulder, M., Wieringa, P. A., Dankelman, J., and Mulder, J. A. 2009. Relationships between driving simulator performance and driving test results. *Ergonomics 52*, 137–153. https://doi.org/10.1080/00140130802277521.

[144] Winter, J. C. de, van Leeuwen, P. M., and Happee, R. 2012. Advantages and Disadvantages of Driving Simulators: A Discussion. *Proceedings of Measuring Behavior*, 47–50.

[145] Wolf, M. C., Muijselaar, M. M. L., Boonstra, A. M., and Bree, E. H. de. 2018. The relationship between reading and listening comprehension: shared and modality-specific components. *Reading and Writing* 2018, 1747–1767. https://doi.org/10.1007/s11145-018-9924-8.

[146] Wu, J., Ahuja, K., Li, R., Chen, V., and Bigham, J. 2019. ScratchThat: Supporting Command-Agnostic Speech Repair in Voice-Driven Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 3, Issue 2*, 1–17. https://doi.org/10.1145/3328934.

[147] Zargham, N., Pfau, J., Schnackenberg, T., and Malaka, R. 2022. "I Didn't Catch That, But I'll Try My Best": Anticipatory Error Handling in a Voice Controlled Game. *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3491102.3502115.

[148] Zargham, N., Reicherts, L., Bonfert, M., Völkel, S. T., Schöning, J., Malaka, R., and Rogers, Y. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–14. https://doi.org/10.1145/3543829.3543834.

[149] Zifonun, G., Hoffmann, L., and Strecker, B. 1997. *Grammatik der deutschen Sprache*. Walter de Gruyter, Berlin.