universität ulm

User- and Situation-Adaptive Voice Output

DISSERTATION

zur Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für Ingenieurwissenschaften,
Informatik und Psychologie der Universität Ulm

Institut für Nachrichtentechnik
(Abt. Dialogue Systems)

von

Daniela Stier
aus Ulm

Ulm, 2022

Amtierende Dekanin: Prof. Dr. Anke Huckauf

Gutachter: Prof. Dr. Dr.-Ing Wolfgang Minker (Universität Ulm)
Prof. Dr. Ulrich Heid (Universität Hildesheim)
Prof. Dr. Michael McTear (Universität Ulster)

Tag der Promotion: 19.09.2022

Fassung 1. Juni 2023

Danksagung

Die vorliegende Arbeit wäre ohne die Unterstützung durch mein Umfeld nicht möglich gewesen. Auf dem Weg zu ihrer Fertigstellung haben mich die verschiedensten Personen geprägt, begleitet und unterstützt. Mein besonderer Dank gilt dabei meinen Gutachtern, Prof. Dr. Dr.-Ing Wolfgang Minker und Prof. Dr. Ulrich Heid, die mir die Möglichkeit einer solchen Forschungsarbeit boten. Ich danke beiden für ihre inhaltliche und wertvolle Betreuung. Außerdem möchte ich mich herzlichst bei Herrn Prof. Dr. Michael McTear für sein Interesse an meiner Arbeit und die Bereitschaft bedanken, die Begutachtung meiner Thesis zu übernehmen.

Diese Arbeit ist das Ergebnis meiner Forschungsarbeit bei der Mercedes-Benz AG, Sindelfingen. Mein herzlicher Dank geht daher an meine fachlichen Betreuer Patricia Kittel und Matthias Boll. Sie standen mir insbesondere in den intensiven Forschungsphasen mit Rat und Tat zur Seite. Ebenso möchte ich mich herzlich bei meinen Doktoranden-Kollegen Jakob Landesberger, Maria Schmidt und Juliana Miehle bedanken, die mich und diese Arbeit bis zum Schluss begleitet, hinterfragt und unterstützt haben. Ein besonderer Dank gilt dabei auch meinen ehemaligen Kollegen und Studenten, Jan Bubeck, Baybora Gülec, Johannes Krämer, Ekaterina Lazaruk, Katherine Munro, Ellen Sigloch, Lara Sohns und Ekaterina Panfilova, die ich während meiner Zeit bei Daimler betreuen durfte.

Letztendlich möchte ich mich meinem privaten Umfeld widmen. Ich danke meinen Eltern, Geschwistern, Familie und Freunden, dass sie mich in meinen “Tunnel”-Phasen ertragen, verstanden und mir geholfen haben. Sie haben mich wenn nötig aufgebaut, unterstützt und versorgt. Nicht nur deshalb sind sie eine wertvolle Bereicherung.

Abstract

Interaction between humans and machines is becoming increasingly voice-based. There are numerous, different contexts in which human-machine interaction can take place, from smart home control to tutor systems to spoken dialog systems in vehicles. As numerous as these interaction contexts are, they each pose individual requirements that need to be taken into account in their development. In this context, the current trend is moving away from formerly command-based interaction toward natural-language, adaptive systems that flexibly adapt to the respective user in a specific interaction context according to the model of human-human communication. This becomes particularly relevant in interaction contexts where a user performs a secondary task in parallel to a so-called primary task, such as interacting with a spoken dialog system while driving a car. For safety reasons, a driver's attention here has to focus on the primary task of driving, while the voice interaction with a system has to run in parallel. In this context, the execution of a secondary task represents an additional cognitive load on the driver, which may affect his or her driving performance and the associated driving safety. In order to counteract this danger according to the model of successful interpersonal communication, the development of an in-vehicle dialog system should take into account both the individual user and the respective interaction context. In this regard, the concept of linguistic alignment represents a particularly valuable mechanism to design human-machine interaction in a natural, efficient, and intuitive way. In this context, this thesis presents an approach for developing a user- and situation-adaptive strategy in the automotive context, focusing on the syntactic design of voice output with reference to syntax as an elementary component of human language. For this purpose, various user studies addressing language perception and production were conducted in order to exploratively characterize individual user characteristics and influencing factors of driver distraction and user experience with different syntactic forms in voice-based interaction besides driving a vehicle. In particular, the user personality within the framework of the Big Five model was proven to be a reliable tool in this context. Based on these findings, this thesis presents a user- and situation-adaptation strategy focusing on

the syntactic complexity of in-vehicle voice output by aligning the syntactic features of realistic spoken driver language with evidenced user preferences regarding the syntactic design of voice output. An evaluation of this adaptation strategy in the context of a user study in actual road traffic revealed a demonstrably enhanced user experience compared to a non-adaptive standard system. In contrast, an effect on driver distraction by applying the developed strategy could not be proven. The findings and achievements presented in this thesis nevertheless provide a valid basis for the development of intuitive, natural dialog systems in dual-task environments.

Kurzfassung

Die Interaktion zwischen Mensch und Maschine findet zunehmend sprachbasiert statt. Dabei gibt es zahlreiche, verschiedene Kontexte, in denen Mensch-Maschine Interaktion stattfinden kann, von der Smart Home-Steuerung über Tutor-Systeme bis hin zu Sprachdialogsystemen im Fahrzeug. So zahlreich diese Interaktionskontexte sind, stellen sie doch jeweils individuelle Anforderungen, die es in ihrer Entwicklung zu berücksichtigen gilt. Dabei führt der aktuelle Trend weg von der ehemals kommando-basierten Interaktion hin zu natürlichsprachlichen, adaptiven Systemen, die sich nach dem Modell der Mensch-Mensch-Kommunikation flexibel an den jeweiligen Nutzer in einem bestimmten Interaktionskontext anpassen. Dies wird insbesondere in Interaktionskontexten relevant, in denen ein Nutzer parallel zu einer sogenannten Primäraufgabe eine sekundäre Aufgabe ausführt, wie die Interaktion mit einem Sprachdialogsystem während des Autofahrens. Aus Gründen der Sicherheit muss die Aufmerksamkeit eines Fahrers hierbei auf der Primäraufgabe des Fahrens liegen, während die Sprachinteraktion mit einem System parallel ablaufen muss. In diesem Kontext stellt die Ausführung einer Sekundäraufgabe eine zusätzliche kognitive Belastung des Fahrers dar, die seine oder ihre Fahrperformanz und die damit einhergehende Fahrsicherheit beeinträchtigen kann. Um dieser Gefahr nach dem Modell erfolgreicher zwischenmenschlicher Kommunikation entgegen zu wirken, gilt es in der Entwicklung eines fahrzeuggebundenen Dialogsystems sowohl den individuellen Nutzer als auch den jeweiligen Interaktionskontext zu berücksichten. Dabei stellt insbesondere das Konzept des linguistischen Alignment einen wertvollen Mechanismus dar, der es ermöglicht, die Mensch-Maschine Interaktion natürlich, effizient und intuitiv zu gestalten. In diesem Kontext stellt die vorliegende Arbeit einen Ansatz zur Entwicklung einer nutzer- und situations-adaptiven Strategie im Automobilkontext vor und fokussiert sich dabei mit Bezug zur Syntax als elementarer Bestandteil der menschlichen Sprache auf die syntaktische Gestaltung von Sprachausgaben. Für diesen Zweck wurden zunächst verschiedene Nutzerstudien zur Sprachperzeption und Sprachproduktion durchgeführt, um daraus explorativ individuelle Nutzereigenschaften und Einflussfaktoren auf die Fahrerablenkung und Nutzererfahrung mit

verschiedenen syntaktischen Formen in der sprachbasierten Interaktion neben dem Führen eines Fahrzeugs zu charakterisieren. Insbesondere die Nutzerpersönlichkeit im Rahmen des Big Five Modells erwies sich hierbei als zuverlässiges Instrument. Auf diesen Erkenntnissen aufbauend wird in dieser Arbeit durch den Abgleich der syntaktischen Merkmale realitätsnaher, gesprochener Fahrersprache mit nachgewiesenen Nutzerpräferenzen hinsichtlich der syntaktischen Gestaltung von Sprachausgaben eine nutzer- und situations-adaptionsstrategie mit dem Fokus auf die syntaktische Komplexität von Sprachausgaben im Fahrzeug präsentiert. Die Evaluierung der Adaptionsstrategie im Rahmen einer Nutzerstudie im tatsächlichen Straßenverkehr ergab eine nachweislich gesteigerte Nutzererfahrung im Vergleich zu einem nicht-adaptiven Standardsystem. Dagegen konnte ein Effekt auf die Fahrerablenkung durch die Anwendung der entwickelten Strategie nicht belegt werden. Die in dieser Arbeit vorgestellten Beobachtungen und Ergebnissen stellen dennoch eine valide Grundlage für den Weg hin zur Entwicklung intuitiver, natürlicher Dialogsysteme in der Sprachinteraktion als Sekundäraufgabe dar.

Contents

Acronyms

ASR	Automatic Speech Recognition
AUT	Autonomous driving mode (SAE Level 5)
BFI	Big Five Inventory (*e.g.*, John *et al.*, 1991; German version by Rammstedt and Danner, 2016)
C	City
CAN	Controler Area Network
CI	Confidence Interval
COP	Comfort Programs
DALI	Driving Activity Load Index (Pauzié, 2008)
DAM	Disambiguation Module
DAS	Driving Assistants
DH	Domain Handler
DM	Dialog Manager
FCV	Final Clause Variant (syntactic realization using a subordinate final clause)
H	Highway
HMI	Human-Machine Interaction
HU	Head-Unit
IQR	Interquartile range
JS	JavaScript
MAN	Manual driving mode (SAE Level 0)
M	Mean
MCV	Main Clause Variant (syntactic realization using two separate main clauses)
Mdn	Median

NCV	Nominalized Clause Variant (syntactic realization using coordinate main clauses with nominalized verbs)
NLG	Natural Language Generation
NLU	Natural Language Understanding
OR	Odds Ratio
QAS	Question-Answer Sequence
RCV	Relative Clause Variant (syntactic realization using a subordinate subject-oriented relative clause)
RH	Research Hypothesis
RQ	Research Question
RTDD	Real-Time Driving Data
SAM	Syntactic Analysis Module
SASSI	Subjective Assessment of Speech System Interfaces (Hone and Graham, 2000)
SD	Standard deviation
SDS	Spoken Dialog System
TTS	Text-to-Speech Synthesis
UC	User Cluster
UEQ	User Experience Questionnaire
WoZ	Wizard-of-Oz
WS	Websocket
XHR	XMLHttpRequest

List of Figures

List of Tables

Chapter 1

Introduction

The interaction between human and machine is increasingly taking place on the basis of speech. Thereby, the trend is moving away from the former command-based style towards increasingly natural, intuitive dialogs based on the human model. One prerequisite for such a development is the ability of a speech dialog system (SDS) to flexibly respond to individual requirements. On the one hand, these requirements relate to the user her or himself: Different users come with different profiles, such as different character traits or, for example, different experiences in interacting with an SDS. On the other hand, these requirements also relate to the context and situation in which the user interacts with the SDS. An intelligent dialog system is therefore expected to adapt to certain user characteristics and preferences in order to make the interaction as efficient and natural as possible.

There are numerous, different and diverse contexts in which human-machine interaction (HMI) takes place, from smart home control to tutor systems. What all these interaction contexts have in common is that a user can focus at 100% on the interaction with the voice assistant or dialog system. This circumstance does not hold when the previously primary task becomes a secondary one, because an additional activity is added to the interaction context and becomes prioritized. An example of such a dual-task scenario can already be found in everyday life, such as cooking, where a voice assistant can assist the cook by going through a recipe: When food is sizzling in several pots and pans at the same time and starting to burn, it seems quite plausible that the focus of the cook in such a situation is more on handling the pots and the voice assistant's stoically recited instructions fade into the background. Another example from everyday life can be found in driving a car in road traffic: While driving in tight city traffic, the focus of an SDS user is primarily on avoiding an accident. In such a case,

the voice assistant should act in a supportive manner and not as an additional distraction. Although the interaction with the voice assistants in the above examples probably takes place differently and different topics are dealt with, both scenarios are linked by the fact that speech must be processed by users in parallel with the prioritized primary task and should not be distracting.

While a user will automatically adapt his or her linguistic behavior according to the cognitive load induced by the external influences of the interaction context, be it cooking or driving (*e.g.,* Kubose *et al.*, 2006; Becic *et al.*, 2010; Vogels *et al.*, 2018), the voice output of an SDS can represent a decisive factor, which is directly perceivable by a user, whether the communication as secondary task has a supportive or rather a disruptive effect on the primary task. Several studies have demonstrated that even the syntactic design of voice prompts can have an influence in this regard (*e.g.*, Demberg *et al.*, 2013, 2016). In this context, interpersonal communication represents a suitable model as a baseline: Human interlocutors adapt their linguistic style to each other intuitively in order to communicate efficiently (Pickering and Garrod, 2004). The so-called principle of *alignment* can thereby be found on all linguistic levels, from lexis to syntax. An obvious example can be imagined in the interaction between a mother and her child or her boss. While the communication in the first case will probably be suitable for children by a comparably simple choice of words and sentence structure, the linguistic style in a professional context will differ and will be adapted according to the environment. In order to enable the most efficient HMI possible according to this model, an SDS should flexibly adapt in its voice output according to the user and the interaction context.

Against this background, the present research work addresses the goal of developing and evaluating a user- and situation-adaptive strategy for the voice output of an SDS in a dual-task environment. Since spoken language is a highly complex and sophisticated human ability (*e.g.*, Moore, 2017; Winograd, 1972), it is neither the intention nor within the scope of this work to address all aspects of language. For this reason, this research work will focus on the syntactic level as one essential linguistic component and elaborate on the syntactic design and syntactic complexity of SDS voice output. For this purpose, the following section will emphasize the effect of varying complexity of different syntactic forms in voice output and present the working hypotheses underlying this work.

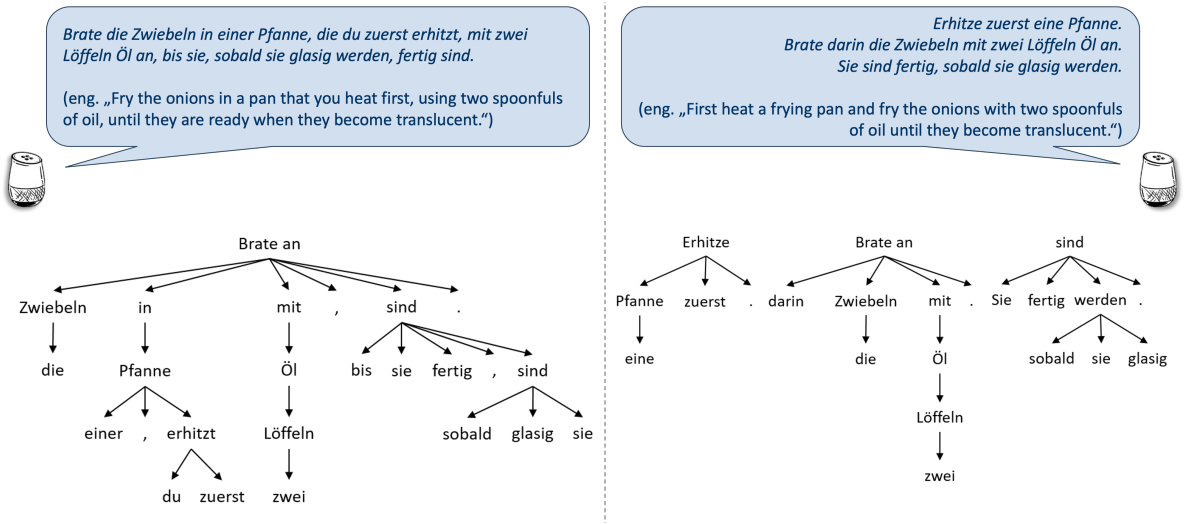


Figure 1.1: Comparison of syntactic paraphrases.

1.1 Syntactic Complexity in Voice Output

An intention can generally be formulated linguistically in an almost infinite number of ways. For instance, the two syntactic paraphrases in Figure 1.1 convey the same content from the interaction context of cooking, but differ in their syntactic structure and thereby concomitant complexity. Generally, a nested relative clause is considered syntactically more complex and difficult to process compared with two linearly organized main clauses with the same content (Warren and Gibson, 2002). The dependency trees provided in Figure 1.1 demonstrate this graphically: Due to the various inserted and partially nested subordinate clauses, the left tree contains many branches and consists of several substructures. In contrast, the variant to the right is represented by a total of three trees resulting from the individual main clauses. This provides a more linear, even structure that appears much clearer and thus easier to understand. This graphical representation provides an organizational overview of the syntactic complexity of the two paraphrases. However, this graphical support is missing at the auditory level. It is thus assumed that the syntactic form in which voice output is designed and the thereby produced inherent complexity can have a major impact on its perception and the interaction with an SDS in general.

On this basis, an underlying Working Hypothesis 1 has been formulated that complex syn-

tactic forms, such as in the left example of Figure 1.1, can increase a user's cognitive load because he or she cannot easily process them and resolve subordinate clauses, for instance. In this way, the design of voice prompts can have a direct influence on how SDS voice output is perceived. The extent of this influence becomes even more concrete in situations where the SDS interaction is deprioritized to the secondary task. Here, the term 'cognitive load' represents a key concept which reveals the necessity of an effective, intuitive user interface without distracting the individual user from his or her primary task. The following section introduces the relevance of a particular dual-task environment.

Working Hypothesis 1:
The complexity of syntactic forms affects a user's perception and experience of SDS voice output.

1.2 Voice Output in a Dual-Task Environment

When interacting with a voice assistant, as in the cooking and driving examples above, the focus is on performing a primary task, while the voice-based interaction has to run in parallel. Despite this similarity, the two mentioned interaction contexts differ in one relevant factor, which is the safety aspect: If the secondary task interferes with the prioritized primary task and demands a distracting level of attention from the user, for example, typically the worst possible scenario is burnt food when cooking, whereas this can result in life-threatening consequences when driving. In this context, the concepts of cognitive load and driver distraction play a central role. It is precisely these life-influencing factors that make the automotive context a generally interesting field of research in the development of SDS concepts for interaction as a secondary task. Here, numerous studies have shown that speech-based interaction while driving is a comparably faster and safer alternative to the visual-haptic control of user interfaces (*e.g.*, Barón and Green, 2006; Weng *et al.*, 2016). However, a similarly large number of studies proved the cognitive load of a driver induced by voice-based interaction and indicated that speech interferes with driving performance (*e.g.*, Nunes and Recarte, 2002; Just *et al.*, 2008; Strayer *et al.*, 2016). In this regard, the acts of listening and talking appear to be equally disruptive (Bock *et al.*, 2007). As outlined above, research exists demonstrating that here a driver intuitively adapts own linguistic behavior according to his or her cognitive

abilities (*e.g.,* Kubose *et al.*, 2006; Becic *et al.*, 2010; Vogels *et al.*, 2018). Similarly, an SDS should contribute to driver safety by taking the needs of an individual driver in a given driving situation into account and by adapting its voice output accordingly. Following the expected effect of syntactic complexity in general, an underlying Working Hypothesis 2 was formulated, indicating that the syntactic form and inherent complexity of voice output in SDS interaction as a secondary task is expected to influence a user's cognitive load and his or her performance of the primary task.

Provided the relevance of the safety aspect, this research will focus on the interaction context of driving as one possible dual-task scenario. Here, the aim is to enhance the user experience and reduce driver distraction induced by cognitive load by adaptively providing voice output in terms of its syntactic complexity.

Working Hypothesis 2:
The complexity of syntactic forms in SDS voice output affects a user's cognitive load and thereby influences the performance of a primary task.

1.3 Thesis Contributions

The goal of this thesis is to define, realize, and evaluate a user- and situation-adaptive strategy for SDS voice output concerning its syntactic form within a dual-task environment like driving. Taking into account the needs of an individual user in a specific driving situation, the aim is thus to enable a natural and intuitive SDS interaction in the vehicle. Compared to a non-adaptive system, this work is intended to both reduce the cognitive load induced by voice-based interaction and improve the user experience. In this respect, this research work addresses the challenge of increasing driver safety in road traffic while at the same time contributing to the long-term goal of developing intuitive, conversational SDSs.

The goal of this thesis is to be pursued from a user perspective. For the development of a syntactic adaptation strategy it is therefore necessary to investigate which linguistic preferences a user has with respect to an SDS and which linguistic properties of a user serve to be adapted by the SDS. For the purpose of these investigations, a suitable data basis is necessary in order to reliably characterize the language perception and production of real users in the vehicle. In summary, the following research steps can be outlined.

1. **User Studies on Language Perception**
   In order to answer the question of how the voice output of an SDS is preferred to be syntactically designed from a user's point of view, it is necessary to collect user preferences regarding different syntactic forms. The goal of this research step is thus to first define and validate approaches which are required as a basis to conduct a user study to gather real user preferences of voice output while driving. This user study will provide insights into the perception of different syntactic forms in voice prompts and factors influencing their perceived naturalness and comprehensibility.

2. **User Studies on Language Production**
   For the syntactic adaptation of voice output, it is necessary to determine to which characteristics an adaptive SDS should adapt. An analysis of the linguistic behavior of real users while driving can serve as a basis for this purpose. The goal of this research step is therefore to build a corresponding corpus of spontaneously spoken driver speech and to investigate whether and which syntactic complexity features allow to characterize an individual user and the particular driving condition.

3. **Development of an Adaptation Strategy**
   In this last research step, the findings of the prior user studies concerning language perception and production are combined to derive a syntactic adaptation strategy for in-vehicle voice output. In addition to providing the theoretical basis, the goal of this research step is to realize the developed strategy in the context of a prototypical implementation and to evaluate it in a user study. This final user study will provide evidence as to whether the working hypotheses defined for this research work can be confirmed.

1.4 Outline of the Thesis

A short introduction into the topic of syntactic complexity in voice output in a dual-task environment was provided at the beginning of this chapter, followed by an overview of the contributions and challenges of this research work. The reminder of this thesis is structured as follows:

Chapter 2 provides the fundamental background and related research for the subsequent chapters. For this purpose, Section 2.1 describes the fundamentals of an SDS. Section 2.2 then introduces the concept of linguistic alignment and syntactic complexity in more detail.

Subsequently, Section 2.3 summarizes the background on cognitive load and driver distraction, before two instruments to model a user's characteristics are described in Section 2.4. Finally, a summary and discussion of challenges of this work are provided in Section 2.5.

Chapter 3 describes the research work performed on the aspect of in-vehicle language perception. First, an approach to manually create syntactic paraphrases is presented in Section 3.1, followed by two user studies investigating the role of syntactic forms in voice ouput while driving and factors influencing their perception in Section 3.2 and Section 3.3. A summary including the fundamental approaches introduced in this chapter and implications on the following research is provided in Section 3.4.

Chapter 4 presents the research work concerning the aspect of in-vehicle language production. In this context, Section 4.1 describes a data collection study for the creation of a spoken language corpus, on the basis of which syntactic complexity components were identified (Section 4.2). Subsequently, Section 4.3 provides the results of the analysis concerning syntactic complexity under consideration of user personality and driving situation. In Section 4.4 the main findings of this chapter are summarized, including implications on the following research.

Chapter 5 describes the development of a user- and situation-adaptive strategy for voice output with regards to syntactic complexity. Here, Section 5.1 presents the details of the development approach and strategy deduction, followed by a description of the prototypical realization to implement the derived adaptation strategy in Section 5.2. Section 5.3 describes the evaluation of the adaptation strategy in a real-life user study. The research work and main findings are then summarized in Section 5.4.

Chapter 6 draws conclusions about the research work presented in this thesis. First, an overall summary including the research contributions is provided in Section 6.1. Finally, suggestions for future work are listed in Section 6.2.

Chapter 2

Background and Related Research

This thesis aims to develop a syntactic adaptation strategy for SDSs under consideration of an individual user's needs in a particular dual-task scenario. In order to account for both these aspects in SDS interaction, a general understanding of the individual SDS components is required. As introduced in Section 1.2, the dual-task environment of driving has been exemplarily chosen for this context and will be focused in the following. Thus, it is furthermore necessary to understand the difficulties and challenges arising from this interaction context as well as the linguistic fundamentals in commmunication that form in-vehicle interaction. For this purpose, this chapter describes the technical and theoretical background for this work. Additionally, related research is introduced and discussed.

The remainder of this chapter is structured as follows: In Section 2.1, the fundamentals of SDSs are described, including an overview of the role of adaptivity and evaluation of SDSs. Section 2.2 introduces the linguistic aspects of alignment and syntactic complexity, which are required as a basis to understand voice-based interaction and to develop an adaptive voice output strategy. Furthermore, a common understanding of syntactic complexity is approached. Section 2.3 then provides an overview of the concept driver distraction and assessment measures. Finally, Section 2.4 describes two instruments for user modelling applied in this work, before a summary and discussion is provided in Section 2.5.

2.1 Overview of Spoken Dialog Systems

The advantages of spoken language, such as linguistic flexibility (Allen *et al.*, 2001) or speed in solving task-specific problems (Cohen, 1992), turn it into a powerful modality in the con-

text of communication understood as information exchange between humans and machines (Fellbaum, 2012). Spoken language thus represents a fundamental component of a spoken dialog system (SDS; Skantze, 2007). Due to the current advances in computer and speech technologies, various prominent working systems have been developed and introduced into commerce, such as the first in-vehicle SDS Linguatronic by Mercedes-Benz in 1996 (Heisterkamp, 2001) or Apple's Siri in 2011 (Pieraccini, 2012), ranging from initially command-based to more conversational systems (Skantze, 2007). Although this binary distinction may fall short in characterizing each SDS, it provides a main distinction of dialog system types: One main characteristic of command-based systems is their limitation in dialog and task complexity (Allen *et al.*, 2001; Fellbaum, 2012; McTear, 2002; Moore, 2017). For instance, a user may utter specific voice commands, which are performed by the system, without any further dialog interaction (McTear, 2002). In this regard, this concept matches the *interface metaphor* proposed by Edlund *et al.* (2008), where an SDS is perceived as a machine interface. In contrast, in a conversational system the interactive aspect is considered essential, which leads to its perception as a conversational partner in accordance with the *human metaphor* (Edlund *et al.*, 2008).[1] Table 2.1 provides a summary of the most salient properties of the two mentioned system types according to Skantze (2007, p. 13).

Following the goal of this thesis to provide natural, intuitive, and efficient interaction in a dual-task scenario according to the human model, the present work ranks among the current research efforts towards conversational dialog systems defined as computer-based applications that enable voice-based interaction as the primary means of communication by means of an interface (Heinroth and Minker, 2012; McTear, 2002).

2.1.1 Fundamentals of Spoken Dialog Systems

As indicated by McTear (2002), SDSs have their origins in the early research of Artificial Intelligence around 1950. However, despite numerous advances in computer and speech technologies, language processing, and dialog modeling since then (Allen *et al.*, 2001; McTear, 2002),

[1] A distinction between the terms 'conversation' and 'dialog' should be considered according to McTear (2004, p. 45). While the term 'conversation' is particularly used to refer to "more advanced dialogue systems that display human-like conversational competencies," the term 'dialog' tends to "signify more restricted systems that engage in specific types of interaction with a more transactional purpose" (*i.e.*, task-oriented). In this thesis, however, both terms will be used interchangeably to refer to computer systems listening to spoken language and using speech to interact with a human.

Table 2.1: Two prototipical classes of SDSs and their associated properties taken from Skantze (2007, Table 2.1).

	Command-based	**Conversational**
Metaphor	Voice interface metaphor.	Human metaphor.
Language	Constrained command-language.	Unconstrained spontaneous language.
Utterance length	Short utterances.	Mixed.
Semantics	Simple semantics. Less context dependence.	Complex semantics. More context dependence.
Syntax	More predictable.	Less predictable.
Language models	Strict grammar, possibly large vocabulary.	Less strict grammar, possibly smaller vocabulary.
Language coverage challenge	How to get the user to understand what could be said.	How to model everything that people say in the domain.

there is a generally applicable architecture for the organization of individual SDS components that interact with each other for successful system functionality (*e.g.*, McTear, 2004; Schmitt and Minker, 2012). An overview of the resulting processing chain, where each subsequent module processes the output of the previous one, is illustrated in Figure 2.1 and comprises automatic speech recognition (ASR), natural language understanding (NLU), dialog management (DM), natural language generation (NLG) and text-to-speech synthesis (TTS).

Following the example provided in Figure 2.1, an SDS user formulates a question like “What’s the weather in Paris?” This utterance is recognized by the ASR component by transforming the speech signal into a textual representation. The following module for language understanding (NLU) interprets this textual basis and converts the user’s intent into a semantic representation. The DM then decides on the next dialog turn. For instance, it determines whether sufficient information from the user is available and communicates with the external application to retrieve the information matching the user’s intent. In the example above, this includes the requested weather information. The DM may also consider to ask for a confirmation or further details by the user, for example, whether the requested location was in France or the United States. The subsequent NLG module takes charge of creating a reponse according to the DM’s decision. The conveyed intention of the DM is converted into words and sentences, before an actual voice output is generated from this textual form by the TTS module. In the

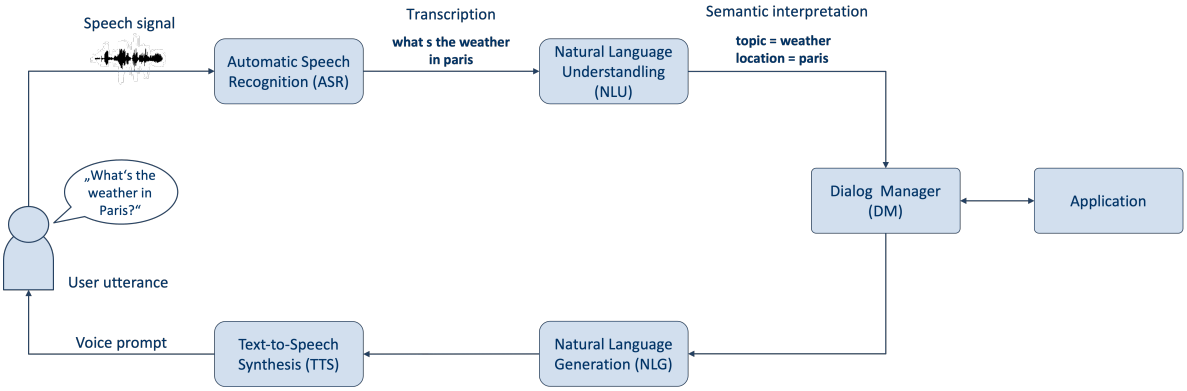


Figure 2.1: Architecture of a standardized spoken dialog system based on McTear (2004).

context of the above example, the SDS may respond to the user's request by generating the sentence "The weather in Paris, France, is sunny with 27 °C."

The individual SDS components are described in more detail in the following.

2.1.1.1 Automatic Speech Recognition

The goal of the ASR module is to convert an acoustic signal into possible recognition hypotheses (Jurafsky and Martin, 2014). For this purpose, the acoustic signal is processed including the removal of noises and channel distortions, and converted into feature vectors to build acoustic models (Yu and Deng, 2016). In combination with a language model, which estimates the probability of an hypothesized word sequence, the textual representation which most likely coincides with the original utterance of the user is returned as a recognition result (López-Cózar *et al.*, 2014; Rabiner and Juang, 1993).

Despite a noticeable progress in the performance of ASR technologies, Young *et al.* (2013) indicated that 15-30% of user requests in many real-world scenarios still result in inaccurate ASR output. In this context, a number of factors may influence the accuracy of the ASR component, for instance, acoustic similarities between words (*e.g.*, López-Cózar *et al.*, 2014), a user's gender (*e.g.*, Abdulla *et al.*, 2001; Levitan *et al.*, 2016; Vergin *et al.*, 1996), his or her age (*e.g.*, Gordon-Salant and Cole, 2016; Potamianos *et al.*, 1997; Russell and D'Arcy, 2007), his or her accent (*e.g.*, Chen *et al.*, 2001; Huang *et al.*, 2004), as well as acoustic distortions

such as disfluencies, hesitations, unknown or mispronounced words, or even ungrammatical, fragmented constructions (*e.g.*, Lamel *et al.*, 2000; Weng *et al.*, 2016). Similarly, certain environments have been proven to represent a challenge for an accurate ASR result, such as public spaces or inside a vehicle (Young *et al.*, 2013). In particular in in-car communication, environmental noise originating from different sources, such as the engine, outside noises, and passenger interaction may impose a detrimental effect on the success of the ASR module (*e.g.*, Cavedon *et al.*, 2005; Chen *et al.*, 2010; Weng *et al.*, 2016).

2.1.1.2 Natural Language Understanding

The output of the ASR component is transmitted to the NLU module to perform a semantic analysis of the user's intent. The aim thereby is to produce a meaning representation of the ASR result by extracting semantic information (McTear, 2002; Strauss and Minker, 2010). The semantic representation of a user utterance is usually recorded in frames consisting of a number of so-called slots (Allen, 1995; López-Cózar *et al.*, 2014). Typically, there are two kinds of semantic representations, including the utterance level defining the user's intent and the word level, which relates to the extraction of information such as named entity recognition (Chen *et al.*, 2017).

Similar to the ASR module, the success of the NLU component is highly dependent on the quality of user input. According to (Cavedon *et al.*, 2005), users tend to produce disfluent, repetitive, and ungrammatical utterances, in particular when they experience cognitive overload. Provided the characteristics of natural language, they also may produce anaphora, ellipses, and ambiguities (López-Cózar *et al.*, 2014). In addition to errors originating from the ASR module, it is thus crucial for the NLU to deal with this type of input data.

2.1.1.3 Dialog Management

The purpose of the DM is to monitor the conversation flow and to coordinate the interaction between a user and the SDS. In this regard, it is responsible to determine a subsequent dialog action in dependence of a user's input and the current dialog state (McTear, 2002, 2004). As such, the DM determines whether sufficient information is available from the input of a user to retrieve and provide the requested information from the external application (Cohen *et al.*, 2004; López-Cózar *et al.*, 2014).

As summarized by Chen *et al.* (2017), the DM module of task-oriented SDSs mainly consists of the two stages dialog state tracking and dialog policy. While the former compares the semantic input information for each turn with the dialog history to manage the current dialog state, the latter defines the next action to be taken according to the current dialog state.

2.1.1.4 Natural Language Generation

The NLG component is responsible to convert the abstract dialog action defined by the DM policy into a natural language utterances (Chen *et al.*, 2017; López-Cózar *et al.*, 2014). Typically, the NLG module consists of three levels (Rambow *et al.*, 2001; Reiter, 1994): In a first step, the content and discourse structure of a system response is defined by the text planner. Subsequently, the structure grammatical relationships to present these contents is defined by the sentence planner, including the selection of lexical items, generation of referring expressions and building of clauses and sentences. Finally, a grammatical response is generated by the surface realizer.

One of the simplest NLG approaches is the template-based approach, which is employed by many SDS (López-Cózar *et al.*, 2014). It presupposes a direct mapping from non-linguistic input by the content planner to a linguistic surface structure (*i.e.*, no syntactic representation is generated; Deemter *et al.*, 2005; Reiter and Dale, 1997). In this regard, a template represents a linguistic structure including gaps, which are to be filled with information provided by the DM (López-Cózar *et al.*, 2014). Although template-based NLG is known for its robustness, an SDS employing this NLG approach may appear tedious due to its repetitiveness in case its templates do not consider a particular degree of variety (Wen *et al.*, 2015a). In addition, this approach is difficult to maintain not easily applied in open-domain systems (Young *et al.*, 2013). To overcome these disadvantages, a trainable generator approach is pursued in current research (*e.g.*, Lemon, 2008; Oh and Rudnicky, 2000; Mairesse and Young, 2014; Stent *et al.*, 2004; Wen *et al.*, 2015b).

2.1.1.5 Text-to-Speech Synthesis

The task of the TTS module consists of converting the textual output of the NLG into speech (López-Cózar *et al.*, 2014). For this purpose, in early SDSs pre-recorded canned speech was

employed for voice output. However, in order to allow for more variation, current systems employ TTS engines to synthesize any arbitrary voice prompt (Cohen *et al.*, 2004; McTear, 2002). Thereby, TTS traditionally comprises the two stages of text analysis and speech generation (López-Cózar *et al.*, 2014; McTear, 2004). While the former transforms the textual basis into a linguistic representation by determining the phonemic structure in words and the underlying composition of the text (Klatt, 1987), the latter actually produces synthetic speech by adding prosodic markers, such as pitch and intonation, and constructing the speech waveform.

According to Cohen *et al.* (2004), several requirements need to be considered in TTS, including intelligibility, naturalness, accuracy, and listenability.

Research is ongoing for all of the described SDS components. Especially the concept of adaptivity has become an popular area in recent research. Current SDSs are occasionally considered to be inflexible as they do not adapt according to a user or dialog flow. Many of them are developed for a stereotyped user in mind (Fischer, 2001), who in reality may rarely exist (Hjalmarsson, 2005a). In this context, the systems' lack of adaptability can prevent a successful interaction and thus leads to an increasing user dissatisfaction (Berg, 2013; Schmitt and Minker, 2012; Ultes *et al.*, 2015). The goal of developing future SDSs is therefore to enable more natural communication by allowing the system to adapt to a user's abilities and needs, and thus resembling human-human communication. The aspect of adaptivity of SDS swill be examined in the following section.

2.1.2 Adaptivity in Spoken Dialog Systems

In the context of human-human communication, it has been observed that human interlocutors continuously adapt to the requirements of the conversation situation and the conversational partner in terms of emotional, acoustic as well as linguistic characteristics (Bell, 2003; Pickering and Garrod, 2004). For instance, depending on the task or the interlocutor, the employed grammatical constructions vary among others (Levelt and Kelter, 1982). In short, humans adapt their conversational strategies according to different factors in order to successfully communicate (Bell, 2003).

Following the human model, SDS users are likewise affected by the linguistic choices in the interaction with an SDS as counterpart. Adaptation mechanisms in the context of human-human communication thus provide important insights for the development and improvement

of adaptive SDSs (Bell, 2003). According to Wärnestål and Kronlid (2014), the overall goal of an adaptive SDS is to adapt to the individual needs of users in terms of their knowledge, preferences, abilities, and objectives, under consideration of the current situation of usage. The task of an adaptive SDS is therefore to adapt to the linguistic choices of an individual user, taking these criteria into account. To achieve this, however, an adaptive system first requires some knowledge about the user. In this context, Papangelis *et al.* (2013, p. 29) define adaptive SDSs as "systems that are able to interact with their users in a more natural and intuitive way than traditional systems/interfaces." In this way, this definition relates with Jokinen (2003)'s remarks, according to which SDSs should be considered as systems that learn dynamically by interacting with humans and can adaptively respond to the user on the basis of individual user models (Jokinen *et al.*, 2004; McTear, 1993). Adaptivity in SDS can be thus be interpreted as the ability of an SDS to accustom different situations and users in order to provide the most efficient form of interaction (Jokinen *et al.*, 2002).

Fischer (2001) and McTear (2004) differentiate *adaptable* and *adaptive* interfaces. The first type empowers an SDS user to personalize the system, for example, by selecting settings, providing feedback to the system and indicating where problems in an interaction occured. Although adaptable SDSs were found to outperform non-adaptable systems (*e.g.*, Litman and Pan, 1999), they are generally considered to provide a rather unnatural way of interaction. In contrast, the second type corresponds to the above understanding of adaptivity, where the system automatically responds to a dynamically changing interaction context. For this reason, this thesis will focus on *adaptive* SDSs in the following.

Schmitt and Minker (2012) proposed the two steps *detection* and *action* to be considered in the development of an adaptive SDS. Here, the *detection* step focuses on the characteristics that can be extracted for adaptation (*i.e.* "what" to adapt). Following these authors, three categories can be differentiated in this context, including interaction-related properties (*e.g.*, user satisfaction, interaction quality, ASR performance), dynamic user characteristics (*e.g.*, emotional state, intoxication) and static user properties (*e.g.*, expertise, age, gender, and preferences). Subsequently, the *action* step subsumes the employed techniques to adapt to the above mentioned features (*i.e.* "how" to adapt). Thereby, the adaptation can be performed on one of three levels consisting of speech input, speech output and dialog strategy. Against the background of this thesis, the following will focus on the second level of speech output.

The previous sections presented fundamentals of SDSs to build a common understanding of the main concepts required for the development of an adaptive SDS. In order to assess

whether the developed system meets the required goals and to verify the working hypotheses underlying this research work, it needs to be evaluated. For this purpose, the following section presents an overview of the evaluation methods applied in this thesis.

2.1.3 Evaluation of Spoken Dialog Systems

Following the usability engineering lifecycle (Möller, 2017), evaluation procedures represent a fundamental aspect in the development of an SDS. One nowadays popular technique to, for instance, analyze user behavior in a particular environment is the *Wizard-of-Oz* (WoZ) approach. In this process, individual system functionalities are replaced by a human-controlled software environment (Grothkopp *et al.*, 2001). In this way, certain functions can already be examined with regard to their usability in advance of a complete system development (Fraser and Gilbert, 1991). Thus, it is possible to test a concept even though the system to be developed has not yet been realized. Following the definition by Bernsen *et al.* (2012), the WoZ approach represents an experimental prototyping method, where the experimenter (the 'wizard') simulates the actions of the SDS to be developed and interacts with the participant. Thereby, the participant believes to interact with a real system. Maintaining this belief is especially relevant because users tend to act linguistically differently depending on whether they are communicating with a human or a machine (Bernsen *et al.*, 2012). It is thus essential to not inform a participant beforehand about the assessment procedure in order to collect natural and unbiased feedback.

The WoZ procedure has been chosen in the context of this work in order to allow for the examination of user behavior in the dual-task scenario of driving. Provided the exploratory character of investigations concerning language perception and production as outlined in Section 1.3, it is considered a valuable instrument. In this context, the usability of developed SDS concepts is assessed by means of subjective evaluation measures in this research work. As compared to objective evaluation, which focuses on system and interaction performances, subjective evaluation deals with the assessment of a system from a subjective user perspective (Möller, 2004). In order to pursue the research goals of this thesis, a controllable frame of SDS interactions needs to be defined. Details and the definition of the interaction scope will be provided starting from Section 3.1.1. For this reason, the objective assessment by means of interaction parameters is hardly applicable and not within the focus of this work. Instead, subjective assessment measures are employed in this thesis by means of surveys and ques-

tionnaires in order to gain insights into the perceived usability and experience of concepts. Due to the explorative nature of this work regarding the initial elaboration of a syntactic adaptivity strategy and its subsequent evaluation, two different approaches were chosen in this context.

2.1.3.1 User Experience of Syntactic Forms in Voice Output

One component of this work is to investigate the extent to which syntactic forms in voice output and their inherent complexity affect the user experience of an SDS user. The syntactic paraphrases created in the course of the work provide the basis for this (s. Section 3.1). For the goal of natural and intuitve voice output, the user studies conducted in this research work rely on the WoZ approach to assess the perceived naturalness and comprehensibility of voice prompts:

- **Comprehensibility** refers to the perceived comprehensibility of voice output. In this regard, this concept asks whether the content of a voice prompt is intuitively and directly understandable.
- **Naturalness** comprises the perceived naturalness of voice output. This concept relates to the question whether a voice prompt is formulated by employing a naturally perceived language style.

2.1.3.2 User Experience Questionnaire

One second component of this work is to investigate the user experience of the developed and realized adaptation strategy focusing on syntactic forms. For this purpose, various questionnaires exist, which focus on user-driven evaluation, such as the Subjective Assessment of Speech System Interfaces (SASSI; Hone and Graham, 2000) or AttrakDiff (Hassenzahl *et al.*, 2003). According to Brüggemeier *et al.* (2020), these questionnaires serve as comparable instruments with regard to measuring user experience. However, for the purpose of this thesis to particularly evaluate user experience of an SDS with a focus on syntactic forms in voice output within a defined interaction scope (s. Section 3.1.1), not all SASSI dimensions and items appeared applicable. As such, the aspects Habitability (*i.e.*, asking about the clarity of interaction as perceived by a user) and Speed (*i.e.*, related to the perceived speed of system interaction) were not considered to contribute to this purpose. However, AttrakDiff and

its focus on hedonic quality, which refers to aspects that are not directly related to complete a goal such as originality of design, were considered as too limited with regard to the expected influence of syntactic forms within the defined interaction scope. For this reason, the comparable User Experience Questionnaire (UEQ) developed for German by Laugwitz *et al.* (2006) was employed as one instrument to assess the usability of the here developed SDS strategy. In contrast to AttrakDiff, the UEQ aims to not only consider hedonic quality aspects but to provide comprehensive insights into subjectively perceived user experience by means of a simple and immediate procedure (Laugwitz *et al.*, 2006). For this purpose, it includes pragmatic and hedonic quality aspects as well as components with regard to the perceived attractiveness Laugwitz *et al.* (2006). In this context, pragmatic quality is focused on goal-oriented aspects, for instance, whether a goal can be achieved efficiently and effectively. In addition, attractiveness is considered as a global rating with regard to approval or disapproval of a system.

Overall, the UEQ comprises 26 items, which consist of complementary adjective pairs. They are subsumed by six dimensions that relate to one of the focused aspects of hedonic and pragmatic quality and attractiveness as global evaluation rating. An overview of the dimension and sample items is provided in Table 2.2.

2.2 Linguistic Aspects in Spoken Dialog Systems

Conducting a dialogue is an inherently collaborative and interactive task (Garrod and Anderson, 1987). According to Clark and Wilkes-Gibbs (1986), the goal of interlocutors is to minimize collaborative effort by phrasing their utterances in a way that allows mutual intelligibility with minimal effort in the shortest amount of time. One element in spoken interaction in this regard is the alignment between interlocutors as a mechanism for achieving this goal (Garrod and Anderson, 1987). The application of alignment in HMI thus represents a reasonable approach to similarly enable successful communication with an SDS (Branigan *et al.*, 2010). Thereby, alignment is found on all linguistic levels, including syntax (Branigan *et al.*, 2003). Against the background of this research work to develop an adaptation strategy with a focus on syntactic forms, in the following, both the aspect of linguistic alignment and syntactic complexity will therefore be introduced.

Table 2.2: UEQ dimensions and quality aspects used in this thesis exemplified by a sample item, based on Laugwitz *et al.* (2006).

Aspect	**Dimension**	**Sample item**	**Item count**
Attractiveness		*attraktiv/unattraktiv* (eng. "attractive/unattractive)	6
Hedonic quality	Novelty	*herkömmlich/neuartig* (eng. "conventional/novel")	4
	Stimulation	*einschläfernd/aktivierend* (eng. "soporific/activating)	4
Pragmatic quality	Dependability	*unberechenbar/vorhersagbar* (eng. "unpredictable/predictable)	4
	Efficiency	*innefizient/effizient* (eng. "efficient/inefficient)	4
	Perspicuity	*verwirrend/übersichtlich* (eng. "confusing/clear")	4
= 3 aspects with 6 dimensions			26

2.2.1 Linguistic Alignment

People therefore tend to adapt their conversational strategies to their counterparts in the course of dialog, and thus exhibit alignment at both non-linguistic and linguistic levels (Branigan *et al.*, 2010; Pickering and Garrod, 2004; Thomas *et al.*, 2018). In this context, alignment is considered to be a mainly unconscious process. At the non-verbal level, for example, alignment involves adjusting to physical behaviors, such as facial expressions (*e.g.*, Bavelas *et al.*, 1986; Navarretta, 2016) and gestures (*e.g.*, Bergmann *et al.*, 2015; Kipp, 2003). They assist an interlocutor in reliably grasping a speaker's goal and intention (Rizzolatti and Fabbri-Destro, 2008). Linguistic alignment, on the other hand, refers to converging linguistic behavior between interlocutors (Branigan *et al.*, 2010). According to the Interactive Alignment Model proposed by Pickering and Garrod (2004), interlocutors align with each other on different linguistic levels. In this process, alignment does not occur isolated on a particular level. Instead, the alignment on one level is interactively enhanced by aligned representations on other levels. The interdependence between linguistic levels has already been demonstrated by several

corpus-based studies (*e.g.*, Wang *et al.*, 2014; Xu and Reitter, 2015). Exemplary cases of linguistic alignment can be found, for example, as phonetic convergence, where speakers align with each other in terms of their pronunciation and accentuation (*e.g.*, Brouwer *et al.*, 2010; Pardo, 2006). This procedure is described as an instantly occuring effect without requiring great cognitive resources (Fowler *et al.*, 2003). Similarly, speakers were found to align in terms of their lexis and semantic-pragmatic choices in order to build a common situation model (*e.g.*, Brennan and Clark, 1996; Danescu-Niculescu-Mizil and Lee, 2011; Garrod and Anderson, 1987). However, not only the content but likewise the form of dialog acts represent a factor influencing the interactive task (Linell, 1998). In this regard, linguistic alignment between interlocutors has also been observed on the level of syntax (*e.g.*, Bock, 1986; Branigan *et al.*, 2000, 2010; Levelt and Kelter, 1982). Similar to the semantic and lexical coordination, interlocutors tend to establish a common syntactic ground (Clark, 1996).

Branigan *et al.* (2010) argue that alignment is based on certain mechanisms that serve as trigger or incentive for its occurence in communication. Instead of focusing on one underlying mechanism, researchers should focus on a combination of them to explain alignment in real-life. A brief overview of such possible mechanisms is provided in the following.

- **Priming:** Alignment is considered as an unmediated mechanism (Branigan *et al.*, 2010). In this regard, it presupposes priming of particular processes and representations, which are not influenced by extralinguistic factors, such as a speaker's belief concerning his or her interlocutor (Pickering and Garrod, 2004). For instance, syntactic priming represents the tendency of a speaker to prefer a particular phrase structure over alternative, available formulations after having used or heard it before (Branigan *et al.*, 2000). The employment of a particular linguistic structure is thus conditioned by its activation, which enhances the probability for further usage.
- **Audience design:** Another mechanism underlying alignment is represented by the concept of interlocutor modelling (Branigan *et al.*, 2010) within the process of audience design (Bell, 1984). It refers to the intuitive assessment of a speaker concerning the appropriateness of a linguistic representation and its application in the interaction with a particular counterpart for the goal of successful communication. According to Clark (1996), a speaker may employ different kinds of evidence to estimate which expression is the most appropriate, such as the cultural background (*e.g.*, cultural group, social position, linguistic competence) and direct interpersonal eperience (*e.g.*, prior interaction). Under consideration of these factors, a speaker may automatically adapt his or her

choice of expressions, for example, by repeating the same syntactic structures or lexical items to avoid miscommunication. Compared to priming, the construction of an interlocutor model represents a cognitively more demanding type of mediated mechanism (Branigan *et al.*, 2010). It is thus assumed that audience design plays a comparably minor role in the alignment between two adult native speakers.

- **Social affect:** Linguistic alignment may serve the goal to establish social relations, for example, by creating positive emotions or expressing affiliation with the interlocutor (*e.g.* Bradac *et al.*, 1988; Van Baaren *et al.*, 2003). In research, this aspect has been evidenced in situations, for instance, where interlocutors differ in terms of their roles (Xu and Reitter, 2015), such as in job interviews or interaction between teacher and students (Jones *et al.*, 1999; Willemyns *et al.*, 1997).

Following the human model, linguistic alignment represents a critical aspect for successful communication. It is thus reasonable that humans transfer their communicative behavior from human-human to human-machine interaction (Branigan *et al.*, 2010). Similarly, alignment in HMI has been observed on all linguistic level, such as phonetics (*e.g.*, Oviatt *et al.*, 2004; Suzuki and Katagiri, 2007), lexis (*e.g.*, Raux *et al.*, 2005; Stoyanchev and Stent, 2009), semantics (*e.g.*, Brennan and Clark, 1996) or syntax (*e.g.*, Branigan *et al.*, 2003; Le Bigot *et al.*, 2007). Branigan *et al.* (2010) indicated that human-machine alignment is even stronger than between human interlocutors to avoid communicative failure. In this context, the degree of convergence appears to be influenced in particular by a user's expectations about the system's expertise. The lower a system's ability to understand spoken language is estimated, the more likely a user is to adapt his or her linguistic behavior to the system (Pearson *et al.*, 2006). In order to overcome this risk of a *habitability* gap arising from a mismatch in technical capabilities (Moore, 2017), computers taking the role of a conversational counterpart need to enable a more sophisticated manner to respond to the requirements of a user in a particular interaction context (Jokinen, 2003). As a consequence, alignment by computers with human users is assumed a promising approach and employed as a general strategy in related research. For example, NLG has gained an increased attention in the area of interactive, adaptive SDSs, which are capable to flexibly adapt voice output under consideration of a particular user to provide the most efficient form of interaction. User-adaptive approaches exist with regards to information presentation (*e.g.*, Lemon, 2008; Moore *et al.*, 2004; Rieser *et al.*, 2010; Walker *et al.*, 2004), sentence planning (*e.g.*, Chen *et al.*, 2002; Hu *et al.*, 2018; Mairesse and Walker, 2007), and surface realization (*e.g.*, Ratnaparkhi, 2000; Varges, 2006; White *et al.*, 2007).

From a user perspective, however, the question remains unsolved whether to apply an adaptation strategy following either a *similarity principle* (*e.g.*, Moon and Nass, 1996; Nass *et al.*, 1995; Thomas *et al.*, 2018) or a *complementarity approach* (*e.g.*, Isbister and Nass, 2000; Lee *et al.*, 2006). While the former indicates to mirror an SDS user's linguistic behavior, the latter suggests to implement a complementary behavior. Furthermore, the exclusive application of one adaptation strategy out of this binary distinction is questionable. Both concepts are found in human-human communication (Dijkstra and Barelds, 2008). It is thus reasonable that also in HMI a mixed approach may be more appropriate. As indicated by Aly and Tapus (2016), the interaction context may contribute to resolve this confusion. Within this research work, the interaction context has been defined as a dual-task environment like driving. To the best of our knowledge, concrete implications for adaptive in-vehicle SDS voice output under consideration of individual user characteristics still are to be defined.

In this section, the concept of alignment has been introduced as one fundamental aspect of successful communication. It has been argued that alignment is perceivable on various linguistic levels in HMI similar to interpersonal interaction. As introduced in Section 1.1, the present research work focuses on the syntactic level and in particular investigates the role of syntactic structures and their inherent complexity in in-vehicle voice output. The following section will thus elaborate on syntactic complexity in more detail.

2.2.2 Syntactic Complexity in Voice Output

From a technical perspective, voice output of arbitrary length and complexity can be generated. However, there is general consent that more intelligent software is required for SDSs to enable complex HMI (Jokinen, 2003). The goal of providing the most efficient form of interaction becomes particularly interesting in situations where SDS interaction is deprioritized to a secondary task, such as in the automotive context. Here, the requirement to support the individual driver and not distract him or her from the primary task, that is, driving, results in the need for efficient and intuitive user interfaces. To this end, the concept of alignment has been introduced as a promising instrument to enable natural, successful communication in HMI. The extent to which an in-vehicle SDS should align with a driver in its syntactic formulations represents the research goal of the present work. The starting point for the investigations in this context is the basic assumption that different syntactic structures exhibit different degrees of complexity and thus represent different levels of cognitive load during their processing.

Syntax is considered to be “vital for language production, as it determines the form of an utterance, which in turn is in a systematic relationship with the meaning of the utterance” (Reitter *et al.*, 2011, p. 1). In this regard, Koch and Oesterreicher (2011) indicate that in particular hypotactical procedures, that is, the subordination in terms of embedding sentence structures within a hierarchically higher sentence frame of a main clause, represent one of the most complex and planning-intensive processes. Nested sentence structures thus contrast with the aggregative character of parataxis, that is, the sequencing of sentences of equal rank, with rather low planning effort and thus an increased possibility for spontaneity. In this context, the production of syntactically complex structures is directly related with an increased cognitive load of a speaker. Thus, the realization of paratactic sentence structures is more likely to be observed in spoken language than complex nesting (Koch, 1995). In parallel, syntax has a significant impact on human language processing, as “[s]ome sentences are harder to process than others, and in some cases this is clearly because of their syntax rather than because of their meaning” (Hudson, 1995, p. 1). Especially in recent years, the subject of the processing complexity of sentences has increasingly become the focus of psycholinguistic research (Bader, 2015). In this context, linguistic complexity is considered as a measure of the cognitive difficulty of human language processing (Liu, 2008). Overall, it has been found that complex sentence structures require a higher degree of a subject’s memory capacity than simpler syntactic structures do (Bartsch, 1973). According to Birkner (2008), relative clauses in particular are among the more complex sentence structures. For example, Kemper *et al.* (2001) demonstrated that the perception of any embedded or subordinate clause is associated with an increased cognitive load. Similarly, Warren and Gibson (2002, p. 79) state the finding that “nested (or center-embedded) syntactic structures are more difficult to process than non-nested structures.” This is directly related to De Saussure (2011)’s principle of linearity in human language: Both the production and processing of a sentence occur linearly, that is, word by word (Liu *et al.*, 2017). German subordinate constructions with verb-last position, which are in a dependency relation with a head, do not allow for linear processing. Accordingly, with increasing distance between syntactically connected words, the processing of a sentence can become more difficult and the human working memory overloaded.

Various measures can be used to objectively assess syntactic complexity in voice output, ranging from surface measures such as sentence length and type-token ratio to a deeper analysis of complexity in the form of syntactic dependencies (*e.g.*, Liu *et al.*, 2017; Pinter *et al.*, 2016). A set of complexity measures will be presented and applied in Section 4.1. In the following, syntactic complexity will be illustrated by means of the example of dependency parse

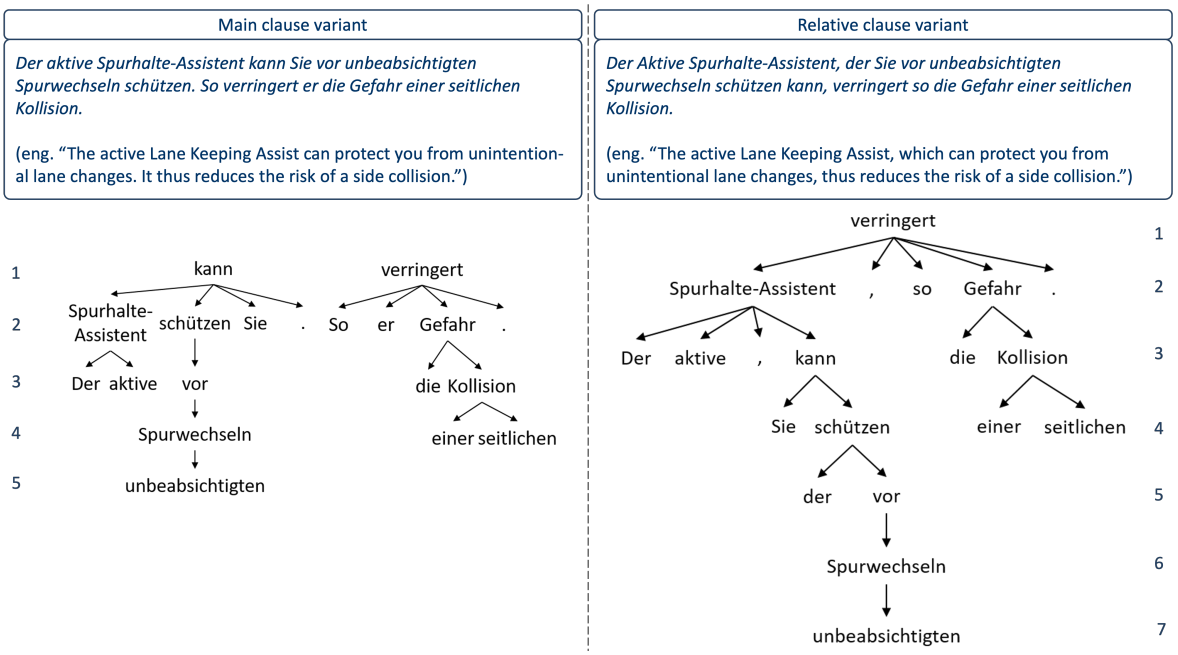


Figure 2.2: The dependency parse tree depth of two syntactic paraphrases exemplifies the complexity of different syntactic forms.

trees. According to Gibson (1998, 2000), this measure represents a valuable instrument as the syntactic complexity of a sentence increases proportional to the length of syntactic dependency. For this purpose, two syntactic paraphrases are compared (s. Figure 2.2). Although both describe the basic functionality of the Lane Keeping Assist, they differ in terms of their syntactic form: While the first one (left) consists of two linearly organized main clauses, the second one (right) contains a nested, subject-oriented relative clause. Moreover, the deduced dependency parses of both paraphrases indicate a depth of five in the case of the main clause variant, whereas the parse tree of the relative clause variant exhibits seven nodes. In accordance with Xu and Reitter (2016), the paraphrase including a relative clause represents a more complex sentence structure compared with the main clause variant, as it exhibits a deeper tree structure.

In the context of spoken interaction in the vehicle, the knowledge of the influence of different syntactic forms and their complexity on a person’s cognitive load is directly related to driver distraction. The following section will therefore introduce the fundamentals of driver distraction.

2.3 Driver Distraction

Nielsen and Minker (2017) pointed out that the adaptability of an SDS towards individual user characteristics and the situational context can help to prevent potential mental overload in the course of interaction with the system. This aspect plays an important role especially in the context of in-vehicle SDSs and a potentially associated driver distraction.

2.3.1 Fundamentals of Driver Distraction

Driving a vehicle on the road is considered as a complex monitoring and control task. In this context, a distinction can be made between primary, secondary and tertiary task types (Kern and Schmidt, 2009): While primary tasks refer to the subtasks required to guide a vehicle, such as speed maintenance and stabilization, secondary tasks include those tasks that contribute to driver safety, such as activating turn signals and windshield wipers. Tertiary tasks are all additional functions that are not necessarily related to the preceding subtasks, such as operating the radio, air conditioning or a dialog system. In this work, tertiary and secondary subtasks are combined analogously to Wierwille (1993)'s definition to subsume those tasks, which should only be performed when the primary task allows to do so.

The execution of a task in the vehicle places certain demands on the driver, which lead to his or her workload (De Waard, 1996). The demands can be of different kinds, such as visual, manual, cognitive and auditory (Wierwille, 1993). As soon as the demands associated with one task become too high, the execution of another task can be impaired and lead to driver distraction (Young *et al.*, 2007). For example, feeling annoyed by the currently played music and thinking to tune to another radio station (*i.e.*, cognitive distraction) as a secondary task besides driving may result in the driver no longer being able to focus his or her attention on the driving task by taking the eyes off the road to the car's infotainment system (*i.e.*, visual distraction) and removing one hand off the steering wheel (*i.e.*, physical distraction) to tune in the desired station. The distraction in this case arises from the division of attention to perform the secondary task in addition to driving the vehicle (i.e., cognitively focus on driving safely, visually checking the road, manually stabilizing the vehicle on the road). Dividing the attention between multiple tasks and their demands in parallel may result as complex and cognitive overload. In the worst case, the driver fails to pay enough attention to the driving task.

In this context, the *multiple resource theory* according to Wickens (2002, 2008, 2020) provides a theoretically based foundation to explain mutual task interference and thus represents a valid framework to understand human attention and cognitive load in dual-task environments. A primary function of the model is the prediction and estimation of interference in simultaneously executed tasks. On the one hand, the complexity of a single task can be determined, but also which tasks interfere with each other (Basil, 2012). The theory fundamentally assumes that people have only limited resources at their disposal. Overall, Wickens distinguishes four categorial dimensions of resources. These are the encoding of information, the modality of perception, the processing level of information, and the manner of response. Accordingly, two tasks interfere more strongly with decreasing efficiency the more common resources are used. Conversely, the more different the resources required during execution, the more efficient is the execution of the individual tasks. The transfer of the model to the primary task of driving can be characterized in particular by the exposure of the visual and auditory dimensions of perception. Here, spatial coordination is particularly relevant, which has to be processed cognitively, such as the position of the vehicle in the course of the road in relation to other road users. A reaction takes place within the framework of vehicle guidance motorically by means of the steering wheel and pedals. Against this background, secondary tasks are critical if they take up the same resources as the primary task. For example, the use of a navigation device interferes with vehicle guidance by conveying spatial information, since the same resources are required for perception (auditory and visual representation), cognitive processing, and reaction execution (motoric handling). In contrast, speech-based communication is usually less disruptive as the auditory channel is engaged with verbal encoding and a linguistic response (Barón and Green, 2006; Vollrath and Totzke, 2003).

However, despite different resources, the perceptual and operational processes of voice-based interaction while driving may nonetheless influence each other (Wickens, 2008): Although, the use of SDSs is generally considered a comparably safe and faster alternative to visual-haptic control of user interfaces (Barón and Green, 2006; Weng *et al.*, 2016), numerous studies have proven the cognitive load of a driver induced by voice-based interaction (*e.g.*, Barón and Green, 2006; Nunes and Recarte, 2002; Strayer *et al.*, 2015a, 2016; Villing, 2009). As such, Strayer and Johnston (2001) demonstrated that speakig on a cell phone while driving led to significant reductions in driving performance. The authors moreover emphasized the difference between a conversation with a conversational partner on the phone or a passenger: While a passenger is, similar to the driver, able to adapt his or her linguistic behavior to the current traffic situation and the driver's cognitive load, this is not valid for a conversational partner

on the phone. Just *et al.* (2008) were able to substantiate these findings with regard to driver distraction caused by voice-based systems and determined that speech processing as a secondary task causes a significant deterioration in driving performance. They concluded that the perception task requires mental resources, which are diverted from the primary driving task performance and thus imposes a negative effect. In constantly changing driving situations, the workload is thereby attributed to the *intuitiveness* and *complexity* of a system (Strayer *et al.*, 2015b). The focus of the development of in-vehicle SDSs is therefore on the lowest possible use of resources and, associated with this, the elimination of potential factors that trigger distraction or attention deficits with regard to vehicle control and thus represent a possible safety risk for the user (Bach *et al.*, 2009).

In this context, Dahlbäck and Jönsson (2007) emphasized the need for the development of situation-adaptive SDSs in the vehicle. With regard to the linguistic design of voice output, they report that drivers in simple driving contexts are generally able to process more complex voice output and exhibit safe driving behavior, whereas this is not the case in complex driving situations.[2] Similarly, Demberg *et al.* (2011, 2013) and Demberg and Sayeed (2011) investigated the design of voice output within the context of information presentation. As such, they proved that syntactically complex voice output in the form of ambiguous subject and object-oriented relative clauses is directly associated with a higher cognitive load than less complex prompts. Overall, the consideration of linguistic complexity in voice output and thus linguistically induced cognitive load by in-vehicle SDSs has been hardly focused on in past research.

Against this background, the extent to which the complexity of voice output in the form of differing syntactic structures cause driver distraction will be specifically investigated in this thesis. For this purpose, the primary driving task represents the focus of voice-based interaction as a secondary task. Thereby, a user-centered approach will be pursued by taking into account the individual needs of an SDS user. However, for the development of a user- and situation-adaptive voice output strategy, it is first necessary to characterize user properties. Therefore, after an introduction to the methodology for the assessment of driver distraction, an overview of the modeling of SDS users will be presented.

[2] Unfortunately, Dahlbäck and Jönsson (2007) only provide a brief summary of their study without any details, for instance, how syntactic complexity was varied. To the best of our knowledge, no further work exists investigating their observations in more detail.

2.3.2 Evaluation of Driver Distraction

Various metrics and methods have been developed to assess driver distraction. According to Green (2001) there are four categories of measurements:

- Primary task performance, *e.g.*, speech control and standard deviation of lane position
- Secondary task performance, *e.g.*, measurement of response times and event detection
- Physiological measures, *e.g.*, heart rate variability and eye movement measurement
- Subjective techniques, *e.g.*, workload ratings

A selection of the most appropriate measurements appears to be difficult. As suggested by Young *et al.* (2009), measures should be chosen under consideration of the competing task. In this work, SDS interaction is investigated under consideration of user experience and driver distraction. For this purpose, several exploratory driving simulation and real-life studies are conducted to investigate the influence of syntactic forms in voice output. Since the scope of SDS interaction considered in this work is limited to a particular domain (s. Section 3.1), secondary task perfomance measures are not applicable. Similarly, as the focus of this work to investigate voice-based interaction, physiological measures such as eye movement are not considered in the following. In contrast, this research work will include, on the one hand, the objective measurement of driving performance and, on the other hand, the subjective evaluation of driver distraction. In the following sections, both the objective and subjective distraction measures will be presented.

2.3.2.1 Objective Measurement of Driver Distraction

Objective driving data can be used to measure driver distraction (Bach *et al.*, 2009). Measurements regarding driving performance are considered to be a reliable indicator of driver distraction and reveal in what way a vehicle is being guided along its intended path (Barón and Green, 2006). Numerous available measures of driving dynamics can be taken into account, such as acceleration, speed, lane keeping quality, distance to the vehicle in front or the standard deviation of the steering wheel angle (*e.g.*, Angkititrakul *et al.*, 2007; Barón and Green, 2006; Young *et al.*, 2007, 2009). Among others, the most frequently employed measures of driver distraction in the context of both simulation studies and real-world driving are objective measures of lateral control (*e.g.*, lane keeping) and longitudinal control (*e.g.*, speed

measures, distance to the vehicle in front) of a vehicle (Bach *et al.*, 2009). In the following, the measures employed in this work are presented:

- **Speed** while driving is commonly used as a measure of distraction, under the assumption that a driver's speed varies to a greater extent when performing a parallel secondary task (Jain and Busso, 2011). Various studies have shown that, for example, that drivers reduce their speed induced by distraction due to concurrent tasks and that there is a higher overall variation in speed (*e.g.*, Horberry *et al.*, 2006; Rakauskas *et al.*, 2004)
- **Distance to a vehicle in front** is employed as another measure of driving performance. Similar to speed, it has been observed that cognitive load and driver distraction is reflected in a weakened vehicle guidance, which is attempted to compensate by increasing the distance between the own vehicle and the one in front (Ranney *et al.*, 2005; Strayer *et al.*, 2003).
- The **lateral lane position** represents an elementary aspect in the context of driving performance. Studies have shown that a driver's lateral position on a lane responds sensitive to the additional performance of secondary tasks (Strayer *et al.*, 2015a). However, different opinions exist in the literature regarding the question whether cognitive load worsens or improves lange keeping (Engström *et al.*, 2017), for example by micro-steering behavior (Li *et al.*, 2018).

2.3.2.2 Subjective Measurement of Driver Distraction

As cognitive load cannot be measured directly, subjective ratings are employed as an additional elicitation method to measure driver distraction. Overall, subjective ratings of driver distraction are considered as essential parameters, which are commonly used in addition to objective measures of driving performance (Pauzié, 2008). Muckler and Seven (1992) consider subjective self-assessments as one of the simplest and appropriate means to measure driver distraction, especially due to its subjectivity, in order to reveal hidden findings in the context of objectively collected measures. In related research, the NASA Task Load Index (NASA-TLX) according to Hart and Staveland (1988) and the Subjective Workload Assessment Technique (SWAT) according to Reid *et al.* (1981) are frequently employed procedures. Both are based on multidemensional scales to account for different workload dimensions (De Waard, 1996). The Driving Acitivity Load Index (DALI) developed by Pauzié (2008) represents a revised ver-

sion of the NASA- TLX, which was specifically adapted to measure distraction in the context of driving tasks and will therefore be applied in this thesis.

A German version was required for the subjective assessment of cognitive load within this research work. For this purpose, the DALI questionnaire employed in the following is based on the version according to Hofmann (2015). In the following, the six DALI dimensions are presented:

- **Auditory demand:** The degree of auditory factors required during the experiment to achieve the overall performance, that is, everything related to listening.
- **Effort of attention:** All the mental (*i.e.*, thinking, deciding, etc.), visual, and auditory factors required in total to achieve overall performance during the experiment.
- **Interference:** Driver distraction and its effect on driving performance induced by the experiment and parallel task completion as a secondary task while driving.
- **Situational stress:** Stress level during the experiment, such as irritation, fatigue, uncertainty, discouragement, etc.
- **Temporal demand:** Perceived pressure and specific impairment due to the sequential nature of tasks during the experiment.
- **Visual demand:** The degree of visual factors required during the experiment to achieve the overall performance, that is, everything related to vision.

2.4 Modelling the User

One crucial aspect for user-adaptive SDSs is the consideration of a user model to enable tracing user characteristics in an interaction context and, consequently, allow tailored system responses accordingly (Hamerich, 2010; Jokinen *et al.*, 2004). According to McTear (1993), a user model is responsible to acquire knowledge about a user and to update it in the course of interactions. In this context, Ryckman (2012, p. 4) defines a user personality as “a dynamic and organized set of characteristics possessed by a person that uniquely influences his or her cognitions, motivations, and behaviours in various situations.” The exploratory nature of this work in identifying potential user-dependent influencing factors in the perception of voice output was pursued by the application of different instruments to characterize user properties in the course of the conducted user studies. They will be described in the following sections.

2.4.1 Technical Affinity Assessment

User experience in in-vehicle SDS interaction forms the focus of this work. In this context, it is reasonable to assume that factors such as prior experiences or general skills in interacting with voice-based systems represent relevant variables to be taken into account (Franke *et al.*, 2019). For instance, a user who is generally open to technology may assess the interaction with an SDS as generally better than a user without this enthusiasm. As one aspect of the exploratory approach in this research work, a questionnaire to measure technical affinity was therefore employed. For this purpose the German Technical Affinity for electronic devices (TA-EG) questionnaire defined by Karrer *et al.* (2009) was included in this thesis. According to the authors, technical affinity represents a personality characteristic, which is represented by a positive attitude, enthusiasm and trust in relation to technical systems.

The TA-EG consists of 19 items, which are subsumed by the four dimensions enthusiasm, competence, positive, and negative attitude.

- **Competence** refers to the self-assessed competence in using a technical device. It is manifested, for example, by knowledge in this area and prior usage experiences.
- **Enthusiasm** This dimension describes the enthusiasm for using technical equipment. It is reflected, for example, in the enjoyment of using them and general interest.
- **Negative attitude** includes aspects related with a negative attitude towards technical devices. This includes, for example, the fear of dependence or their uselessness.
- **Positive attitude** comprises aspects relating to the positive attitude toward technical devices. These include, for example, their role in everyday life and long-term effects.

An overview of the dimensions and sample items is provided in Table 2.3.

2.4.2 The Big Five Personality Model

Human personality has frequently been described in the framework of the Big Five Model, which assumes that human personality can be defined by means of few basic dimensions (Costa and McCrae, 1992, 1999; Goldberg, 1990; John *et al.*, 1991). In psychology, it has become a standard approach to describe a personality due to its robustness and universal nature (McCrae and Costa, 1997) and is often referred to as the OCEAN model. Thereby, the

Table 2.3: Dimensions and example items of the TA-EG questionnaire, based on Karrer *et al.* (2009).

Dimension	Sample item	Item count
Competence	*Es fällt mir leicht, die Bedienung eines elektronischen Geräts zu lernen.* (eng. "I find it easy to learn how to operate an electronic device.")	4
Enthusiasm	*Es macht mir Spaß, ein elektronisches Gerät auszuprobieren.* (eng. "I enjoy trying out an electronic device.")	5
Negative attitude	*Elektronische Geräte verringern den persönlichen Kontakt zwischen den Menschen.* (eng. "Electronic devices reduce the personal contact between people.")	5
Positive attitude	*Elektronische Geräte erleichtern mir den Alltag.* (eng. "Electronic devices make my everyday life easier.")	5
= 4 dimensions		19

acronym represents the five different traits corresponding to the factors **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism. In general, each of these factors can be interpreted as a scale between two poles, such as high and low extraversion.

- **Openness** describes the creativity and imaginativeness of a person (Durupınar *et al.*, 2011). A highly open personality is considered as creative, prone to esthetics and new ideas, and have a "rich and complex emotional life" (Costa and McCrae, 1992, p. 6). In contrast, a person at the other end of the continuum with a low level of openness is described as conforming and conventional (Rammstedt and Danner, 2016).
- **Conscientiousness** refers to a person's tendency to be organized, careful and disciplined (Durupınar *et al.*, 2011; Rammstedt and Danner, 2016). Highly conscientious people are considered to be persistent, meticulous, and efficient, while low conscientiousness is equated with low organizational skills and apathetic attitudes (Costa and McCrae, 1992).
- **Extraversion** comprises a comparably wide range of attributes. As such, highly extraverted persons are considered as outgoing, sociable, and independent, while introverted ones are rather perceived as cautious and reflective (Gill and Oberlander, 2002).

- **Agreeableness** is mainly related to interpersonal behavior (Costa and McCrae, 1992). While an agreeable personality is referred to cooperative, friendly and thoughtful persons, low agreeable individuals are described as insensitive, cynical, and hostile (Durupınar *et al.*, 2011).
- **Neuroticism** describes the emotional stability of a person and his or her tendency to feel stress and negative emotions (Costa and McCrae, 1992; Durupınar *et al.*, 2011). A highly neurotic person is described as anxious, emotionally instable and with a low self-confidence. In contrast, low neuroticism is manifested in emotional stability and self-confidence (Gill and Oberlander, 2002; Rammstedt and Danner, 2016).

A number of studies have investigated the relationship of personality characteristics and behavior (*e.g.*, Durupınar *et al.*, 2011; Paunonen and Ashton, 2001), and in particular with relation to language use (*e.g.*, Fast and Funder, 2008; Gill and Oberlander, 2002; Mairesse and Walker, 2007; Metze *et al.*, 2011). For instance, extraverts have been shown to speak louder with less hesitations compared to introverted persons (Scherer and Scherer, 1981). Similarly, extraverted individuals were observed to be more talkative, produce informal speech and use fewer negations (Burnett and Ditsikas, 2006; Gill and Oberlander, 2002). Furthermore, neurotics were characterized with a low lexical density (Gill and Oberlander, 2002). Against this background, most of recent computational applications to account for human personality appears to focus on individual personality traits in isolation, such as extraversion, which is considered as the easiest factor to model (Mairesse and Walker, 2007), and their extremes (Mairesse and Walker, 2011). However, human personality consists of the different personality factors *simultaneously* on a manifestation range *between* two extreme poles. It is therefore necessary to apply a more fine-grained personality model in order to account for individual differences. The extension of previous approaches will therefore be the focus of this work.

As human personality has been shown to be reflected language behavior, the Big Five Model represents a valuable framework to attribute linguistic differences in spoken language. In this research work, the validation for the German adaptation of the Big Five Inventory (BFI; *e.g.*, John *et al.*, 1991) according to Rammstedt and Danner (2016) is employed. This BFI questionnaire consists of 45 items, which are assigned to the individual Big Five traits. An overview of the personality factors including sample characteristics and an exemplary questionnaire item is provided in Table 2.4.

Table 2.4: Personality traits, facets and exemplary BFI item, based on Rammstedt and Danner (2016).

Factor	Sample facet	Sample item	Item count
Agreeableness	Altruism, accommodation	*Ich bin hilfsbereit und selbstlos gegenüber anderen.* (eng. "I am helpful and altruistic towards others.")	10
Conscientiousness	Neatness, self-discipline	*Ich erledige Aufgaben gründlich.* (eng. "I complete tasks thoroughly.")	9
Extraversion	Assertiveness, activity	*Ich bin gesprächig, unterhalte mich gern.* (eng. "I am talkative, like to chat.")	8
Neuroticism	Anxiety, depression	*Ich bin deprimiert, niedergeschlagen.* (eng. "I am depressed, dejected.")	8
Openness	Openness to esthetics and new ideas	*bin originell, entwickle neue Ideen.* (eng. "I am inventive, I develop new ideas.")	10
= 5 factors			45

2.5 Summary and Discussion

This section first provides a summary of this chapter, before related work is discussed and arising challenges within the scope of this research work are highlighted.

2.5.1 Summary

In this chapter, the technical and theoretical background for this research work was introduced. For this purpose, an overview of SDSs in general was provided in Section 2.1. Here, first the fundamentals concerning the individual SDS components were described (s. Section 2.1.1), before the most relevant concepts for the development of an adaptive SDS were introduced (s. Section 2.1.2). Finally, the background relevant in the evaluation of SDSs was presented in Section 2.1.3. Here, the focus was on the WoZ procedure and UEQ questionnaire, which are applied in this work to evaluate user experience. Subsequently, Section 2.2 provided back-

ground for linguistic aspects in the development of adaptive SDSs. Thereby, Section 2.2.1 argued to pursue the human model and that the application of linguistic alignment in HMI represents a valuable approach to enable the most efficient and successful interaction with an SDS as possible. Likewise in the context of voice-based interaction, Section 2.2.2 substantiated and demonstrated the complexity of differing syntactic forms by means of an example. In the context of voice-based interaction and the effect of syntactic complexity, Section 2.3 provided background on driver attention and distraction. First, the fundamentals were clarified (s. Section 2.3.1). Second, an overview of evaluation methods to assess driver distraction were described with a focus on the metrics applied in this work (s. Section 2.3.2). In Section 2.4, the exploratory nature of this work to develop and evaluate a syntactic adaptation strategy, taking into account individual user characteristics and the interaction context, is reflected. In order to enable the adaptation of an SDS to a particular user, an understanding of user characteristics is required. For this purpose, two instruments were described, which are applied in this work.

In the following, related work and arising challenges are discussed. Here, the research goal of developing and evaluating a user- and situation adaptive voice output strategy for SDS interaction in the dual-task scenario driving is focused.

2.5.2 Related Research and Challenges

The goal of this research work is to deduce an adaptation strategy for SDS voice output in order to increase user experience and decrease cognitive load of a user in SDS interaction as a secondary task. For this purpose, the primary task of driving has been chosen as a possible dual-task environment, as it requires special attention with regard to driver distraction and the related safety aspect. The focus is thereby on the syntactic design of voice output, which is assumed to affect an SDS user's perception, experience and cognitive load. Thus, for the context of this thesis, the interdisciplinary research aspects of linguistics, psychology, and computer science arise.

On the one hand, there are linguistic studies on the complexity of language, and on the psycholinguistic level on language comprehension and production. On a theoretical level, Koch and Oesterreicher (2011) point out that subordinated syntactic structures are probably among the most complex and demanding procedures in language production. Similarly, Hudson (1995) proved that sentences can be difficult to comprehend because of their syntactic form. Warren and Gibson (2002) further explained that nested structures in particular are

more difficult to process than linear, non-nested ones. In sum, syntactic complexity is shown to be directly related to increased cognitive load in language production and processing tasks.

Although these theoretically based observations refer to the aspects of speech production and perception as the sole main task, they can form the basis for transfer to spoken language in HMI and can be referred to concretely in the design and development of conversational SDSs. However, the focus in this respect in the related literature has been more on defining guidelines for interaction design and dialog guidance, taking into account both structural and technical aspects. For instance, Branham and Mukkath Roy (2019) examined guidelines which are taken into account in the development of commercial voice assistants based on the model of human communication, and highlighted, among others, that syntactic complexity in voice output should be avoided. Similarly, based on the analysis of user behavior in WoZ studies, Large *et al.* (2017, 2019) derived guidelines with respect to the design of conversational user interfaces in vehicles. However, little attention has been paid to the concrete application and implementation of these concepts, especially with respect to the linguistic design of voice output and its complexity under consideration of the interaction context. Thus, to the best of our knowledge, only isolated research papers addressed the role of syntactic complexity in in-vehicle SDSs. For example, in this context, Demberg *et al.* (2011, 2013) showed a direct relationship between syntactic forms in the setting of ambiguous subject- and object-oriented relative clauses on a driver's cognitive load and driving performance. Evidence for this was also provided by Dahlbäck and Jönsson (2007), although details on their procedure and the studied complexity aspect are lacking. Overall, the extent to which these findings can be applied in terms of individual user preferences and needs to enhance user experience and reduce a driver's cognitive load is left open in these works. In a recent paper, Meck and Precht (2021) presented a more concrete approach to voice output design by examining, among others, the influence of various syntactic parameters, such as word order, sentence length, and structure, in the context of a user study. While this work takes the aspect into account, that various linguistic parameters can influence the perception of voice output, it is limited to prompts in written form and does not consider the individual user within the interaction context of driving.

While the above works focus on voice-based interaction in the vehicle as an interaction context, but lack consideration of the individual user and his or her requirements, there are, on the other hand, numerous research studies that deal with the language of individuals. In this context, especially the Big Five Model is applied and to what extent linguistic characteristics

are assigned to different personality traits. For example, Mairesse *et al.* (2007) provided a summary of the linguistic properties of the personality trait extraversion and demonstrated the use of more complex syntactic structures by extroverts and simpler syntactic forms by introverts. An example of the realizations of an interactive system can be found in Mairesse and Walker (2010), who developed with their PERSONAGE system a parameterized generator that produces utterances according to different expressions of the Big Five Personality Traits. Although the authors were able to prove the perception of the intended personality traits by the output generated by PERSONAGE, in the context of spoken interaction the discussion remains which preferences a person has regarding the personality of his or her counterpart. In this regard, the work of Thomas *et al.* (2018) evidenced the preference of an interactive system with similar personality traits, while Lee *et al.* (2006) observed the opposite. Therefore, in the present work, it was argued that considering the interaction context can provide information about a decision regarding the adaptation principle, which is missing in the above works.

Overall, this research work aims to combine the here introduced research aspects by accounting for linguistic and syntactic complexity based on the human model of communication and transfer these concepts to the design of voice output in the automotive context. Besides the interaction context of driving, the consideration of individual user characteristics represent a further central component. This thesis therefore aims to connect the two aspects of language production and language perception and to investigate how, from a user's point of view, spoken interaction has to be designed on a syntactic level in order to meet the demands of the parallel primary task of driving a car. A combination of these factors is considered as relevant for the goal to develop a user- and situation-adaptive strategy for syntactic complexity in in-vehicle voice output. The following chapters will introduce the research work, which has been conducted for the purpose of this thesis goal.

Chapter 3

User Studies on Language Perception

As introduced in Chapter 2, this thesis is targeted at conversational dialog systems, which unlike command-based systems build on the principles of human communication. The integration of natural, spontaneous speech in conversational SDSs is intended to translate the advantages of interpersonal communication to HMI in the vehicle. Following the model of linguistic alignment between interlocutors, it is particularly expected to provide an intuitive and efficient way of interacting in dual-task environments. However, the nature of interaction as a secondary task results in special requirements for the voice output of an SDS: It needs to be processed in parallel and should not distract the driver from his primary task of driving. In this respect, an SDS's voice output should be intuitively perceivable and its complexity should take the cognitive capacity of the driver into account. According to the requirements of audible language by Wachtel (2003), this implies a form and an organization of information that is pronounceable and comprehensible for the listener. Similarly, Chomsky (2014, p. 11) refers to linguistic expressions that users regard as *acceptable*, since they were "more likely to be produced, more easily understood, less clumsy, and in some sense more natural." The design of in-vehicle voice output should thus aim at *acceptable* prompts in terms of their complexity and therefore the concepts of the perceived *naturalness* and *comprehensibility* play a central role in the manual preparation and subsequent evaluation of voice output.

Against this background, this chapter focuses on the aspect of language perception in a dual-task environment from a user's perspective. Under consideration of the structural complexity of voice output, the appropriateness of syntactic forms in in-vehicle prompts is investigated in terms of their perceived naturalness and comprehensibility. In this context, first fundamental relations needed to be clarified, for instance, whether the syntactic form of a

voice prompt influences the user's perception at all and to what extent this influence is related to individual characteristics of the user and the driving situation. For this purpose, user studies were conducted on the basis of manually created syntactic paraphrases. Their generation approach was evaluated in an initial pilot study to demonstrate the semantic and syntactic comparability of the purposely created voice prompts in order to provide a reliable framework for the investigation on the influence of syntactic forms. A second pilot study was conducted to investigate the level of consciousness of syntactic forms. The goal here was to investigate whether participants were able to identify and distinguish different syntactic forms in voice output and whether their (lacking) awareness of structures allowed intuitive user ratings on a subjective level. On this basis, two subsequent user studies in a driving simulator were conducted. While the first one aims at demonstrating the general relevance of syntactic forms while driving, the second was designed to further reveal individual user and situation-specific characteristics, which relate to the perception of syntactically differing voice prompts.

In this chapter, first, preliminary work is described in Section 3.1, where a detailed overview of the manual approach to create syntactic paraphrases and its proof of validity are given. Second, Sections 3.2 and 3.3 include a description of the methodology and results of the subsequent user studies on the perception and role of syntactic forms in voice output. Finally, a summary of the obtained observations and results are presented in Section 3.4.

3.1 Manual Preparation of Syntactic Paraphrases

This section describes the process to manually prepare syntactically differing voice prompts as the basis for subsequent investigations. The focus of this approach was to create syntactic paraphrases of a comparable semantic complexity in terms of content and information density to allow conclusions about syntactic differences. Since, to the best of our knowledge, no validated procedure is known for this purpose, a custom approach was developed.

As Moore (2017) emphasized, spoken language is a highly complex and sophisticated human ability. By its nature, it imposes restrictions on currently feasible HMI. Therefore, it is a common approach in the development of an SDS, including conversation-based systems, to limit its capacity to a controllable and clear scope and thereby constrain a user's expectations and behavior. This approach was adopted in terms of the definition limitations and requirements.

Example 3.1: The dialog turns in A and B clearly differ in terms of number and linguistic form. Missing user information in B is requested by the SDS, leads to additional turns and the use of linguistic material without a comparable basis in A.

Turn	Dialog A	Dialog B
1 User:	*Fahr nach Ulm in die Sonnenstraße 13.* "Drive to Ulm to the Sonnenstraße 13."	*Navigiere mich nach Ulm, bitte.* "Navigate to Ulm, please."
2 SDS:		*Wie lautet die Adresse?* "What is the address?"
3 User:		*Das ist die Sonnenstraße 13.* "That is Sonnenstraße 13"
4 SDS:	*Ok, ich starte die Navigation.* "Okay, I'll start navigation."	*Ok, ich starte die Navigation.* "Okay, I'll start navigation."

In this section, first the selected scope is presented, followed by a description of the procedure to manually prepare syntactic paraphrases. Finally, the results of two pilot studies are described concerning the validation of the proposed approach and the level of consciousness of syntactic forms in the perception of participants, before the approach is applied and extended. The work presented in this section is based on the publications by Stier and Sigloch (2019) and Stier *et al.* (2020b).

3.1.1 Definition of Scope

Prior to the preparation of prompts, a number of assumptions was made. These are presented in the following paraphraphs.

Limitation of Dialog Turns. In order to determine acceptable syntactic structures according to Chomsky (2014) in the context of in-vehicle voice output, the scope of interaction between a user and a goal-oriented dialog system was limited to consistent and controllable one-shot sequences, which comprise a user request and a corresponding system reply without any follow-up actions. This restriction on the scale of dialog turns is motivated by the fact that a more extensive dialog with several user or system queries would not be suitable for a comparison at this point, for example, in the case of missing user information (Example 3.1): Both the linguistic form and number of dialog steps would presumably differ over a few users already.

This would not permit a focused analysis and systematic conclusions targeted at syntactic forms. It is therefore essential to standardize the dialog scope by means of the smallest possible unit of a human-machine dialog, that is, one-shot sequences between user and SDS.

Definition of the Prompt Length. In order to explicitly vary syntactic forms, a certain information density is required. A system response like *Ok, ich starte die Navigation* (eng. “Okay, I’ll start navigation”) as shown in Example 3.2, which consists of a single main clause with a maximum of two propositions, provides limited possibilities for syntactic variation. Only a number of propositions and related phrases allows a different linking of information and thus to systematically change syntactic structures. For this reason, a minimum length of two phrases was specified. Only under this requirement the general applicability of syntactic coordination strategies can be guaranteed.

Example 3.2

Voice Prompt	**Propositions**
Ok, ich starte die Navigation.	1- Ok
“Okay, I’ll start navigation.”	2- starte(ich, navigation)

Explanations as Use Case. As introduced above, the comprehensibility of in-vehicle voice output is of essential importance. Against this background, the concept of an explanation represents a particularly suitable context of investigation. An explanation generally aims to convey existing associations or facts. An explanatory system output in response to a user request can thus be used in particular to assess the subectively perceived comprehensibility in dependence of its syntactic form. In the following, one-shot Question-Answer Sequences (QAS) are therefore considered as an appropriate use case for speech-based interaction in vehicles according to the above requirements. We are aware that an explanatory voice prompt according to the above conditions will presumably represent a generally higher cognitive load for a driver than simple, short prompts (*e.g.*, Example 3.2). However, it is assumed that a certain complexity is to be accepted in order to guarantee the purpose of these investigations, that is, the applicability and impact of different syntactic forms in vehicle-related voice output.

3.1.2 Question and Explanation Types

For the targeted preparation of explanatory voice prompts, possible explanation types which may be relevant in the context of one-shot QAS in the vehicle were considered and which

question types may be used to initiate them. The results of these considerations are described in this section.

Roth-Berghofer and Cassens (2005) defined explanations in human-human communication as important means of conveying information. On this basis, the authors referred to Sørmo *et al.* (2004; 2005) and suggested five goals of explanations, which apply for knowledge-based systems in general. Similarly, in the present context of QAS these explanation goals motivate a person's need for an explanation and provide information about its content. In addition to the general purposes of an explanation, its interpretation with reference to the current context is explained below.

- **Transparency:** Explanation how a system reached an answer.
  This type of explanation is based on a desire to explain system behavior and understand how a system's response was found. Applied to the vehicle context, such an explanation could be required, for example, if an existing routing is adjusted unexpectedly due to increased congestion.
- **Justification:** Explanation why an answer is a good answer.
  The lack of explanation of an event can trigger this kind of explanation. Thereby, it can increase the confidence in a system. In the vehicle context, the sudden failure of a driving assistant due to snow-covered sensors could call for such an explanation.
- **Relevance:** Explanation why an answer is relevant.
  An explanation of this kind can clarifiy a system's strategy and the background for a response. Applied to the vehicle context, such an explanation could reveal that the reason for an explicit request to activate the Bluetooth function to connect a mobile device to the vehicle is that this step is often missed.
- **Conceptualization:** Explanation of the meaning of concepts.
  This type of explanation is based on a misunderstanding due to an unclear terminology or concept. In the vehicle context, an explanation of this kind can for example clarify the term "instrument cluster" to a novice driver.
- **Learning:** Explanation to teach a user about a domain.
  In general, this type of explanation is applied to the desire to learn unknown functionalities. The focus here is on a description of the process to solve a problem. Applied to the vehicle context, such an explanation can help where and how to active the hazard warning lights.

Based on the work by Spieker (1991), Roth-Berghofer and Cassens (2005) mapped these explanation goals derived from the user needs outlined above onto different types of explanations. This procedure allowed to establish an association of which goal is met by which type of system explanation given to a user. However, in general one type of explanation can be used to fulfill different goals of explanations:

- **Cognitive explanations** explain and help to understand system behavior. For instance, they can be related to the goal of transparency or justification by answering a question like “How/Why did the system come up with this answer?”
- **Why-explanations** provide the inquiring user with a reason and justification for a system’s behavior. For example, they provide an answer to the question “Why does the system do that?” and can be related to the goal of relevance or justification.
- **How-explanations** are considered as a special type of Why-explanations that describes a process as a causal chain which leads to an event. For instance, they can answer the question “How does this work?” and may fulfill the goal of transparency or learning.
- **Purpose explanations** provide information about the meaning or purpose of an object or thing. As an example, they may answer a question like “What is this for?” and can among others be related with the goal of relevance.
- **Conceptual explanations** enable the understanding of a new concept with the help of an already known concept. For instance, they can provide an definitional answer to the question “What is ...?” or “What is the meaning of...?” and can be related to the misunderstanding of a concept or the goal of giving a theoretical justification.

In the context of the present work, by considering possible question and explanation types, it became obvious that not all kinds of an explanation were suitable for one-shot QAS. In the case of cognitive, purpose and why-explanations, previously experienced system behavior is required. Thus, they did not meet the requirement of one-shot interaction without follow-up actions either by the user or SDS. Similarly, in the case of how-explanations conceptual knowledge is presumed. For this reason, conceptual explanations are focused in the following since they meet the requirements of one-shot QAS without depending on the knowledge perceived from further explanation types. Table 3.2 concretizes conceptual explanations as interpreted in the present work and provides an example.

Table 3.2: Conceptual explanations in the present context of one-shot Question-Answer Sequences, exemplarily demonstrated by means of a driving assistant.

Conceptual explanations	A conceptual explanation follows questions of the form “What is ...?” The aim of this type of explanation is to establish a link between unknown (*e.g.*, an unknown driving assist) and known concepts (*e.g.*, description of the applicational scope). The transfer of factual knowledge is enabled in the form of a definition.
Example question	*Was ist der Brems-Assistent?* (eng. “What is the Brake Assist?”)
Example answer	*Der Brems-Assistent ist ein Fahrassistent. Er bremst Ihr Fahrzeug ab, um Unfälle mit Fahrzeugen sowie Fußgängern zu vermeiden.* (eng. “The Brake Assist is a driving assistant. It brakes your vehicle to avoid accidents with vehicles as well as pedestrians.”)

3.1.3 Requirements for Explanatory Voice Output

After narrowing the scope for QAS to conceptual explanations, this section considers general requirements for the voice prompts to be prepared.

According to Maybury (2004, p. 3), the task of “Question answering (QA) is an interactive human computer process”, which includes “presenting and explaining responses in an effective manner.” Within this context, Griceans maximes (Grice, 1975) represent a commonly employed guideline underlying cooperative conversation. According to this theoretical basis, every dialog or conversation presupposes cooperation between the interlocutors, which is essentially based on four maxims:

- **Quality:** Say only true things
- **Quantity:** Be as informative as required
- **Relation:** Focus on the relevant things
- **Manner:** Be clear and concise

In this work, an SDS assumes the role of the interlocutor, whose responses should be designed according to the above guidelines for successful and effective communication. For this purpose, the maximes were interpreted against the background of QAS in this work.

Quality. An explanatory voice prompt should only contain true information. Thus, it should be based on true knowledge, which is considered as a criterion for voice output.

Quantity. An explanatory voice prompt should be as informative as required. Thus, it should precisely contain the necessary information to answer the question by an SDS user without redundancy. The completeness of information is therefore considered a criterion in the preparation of voice output.

Relation. An explanatory voice prompt should focus on the relevant things. Relevant is considered whatever serves the successful communication, that is, the understanding of an explanation. This is perceived by an adequate information structure, including the theme followed by the rheme. The appropriateness of the requested information in terms of a comprehensible arrangement is therefore another criterion for voice output.

Manner. An explanatory voice prompt should be clear and concise. Thus, it should avoid obscure expressions and ambiguities. The comprehensibility in terms of lexical and syntactic properties is considered as a further criterion for voice output.

The above criteria serve as a guideline for the creation of explanatory voice output embedded in one-shot QAS. Fundamental to the investigation of the acceptability of voice prompts according to Chomsky (2014) is thus the fulfillment of the criteria of quality, quantity, relation and manner according to Grice (1975). Taking these aspects into account, variants of a conceptual explanations were created in the context of one-shot QAS using different syntactic realizations. The approach followed here is described in the following section.

3.1.4 Methodology and Preparation of Syntactic Paraphrases

After defining the scope of the present work on one-shot QAS and conceptual explanations with corresponding question type What, this section describes the approach and the employed method for preparing syntactic paraphrases. The approach presented in the following first involves the selection of suitable information and its structuring in order to achieve semantic-syntactic comparability of conceptual explanations for different topics. Using selected measures, their comparable semantic complexity was additionally aimed at. Starting from this basis, paraphrases were created in the form of different syntactic realizations. First, the approach is demonstrated using the domain of driving assistants and, in a next step, applied and extended to a second domain of comfort functions.

3.1.4.1 Selection of Content: Ensuring the Criteria Quality and Quantity

In the context of QAS in the vehicle, questions and explanations about vehicle functions are a particularly suitable choice. Against the background of Grice's (1975) criterion of quality, instruction manuals provide a high-quality basis for the targeted extraction of information and subsequent processing in the form of voice output. As the flagship of the Mercedes-Benz premium class, digital and print manuals of the S-Class model[1] were therefore used, which is equipped with advanced technologies and functions.

In general, instruction manuals contain explanations of vehicle-relevant functionalities and applications. Since the topic of driving safety is particularly relevant in the vehicle context, driving assistants (DAS) were defined as a first QAS domain. In the following, the method for preparing syntactic paraphrases will be explained on the basis of this domain. Subsequently, the presented approach will be applied to a second domain concerning comfort functions (COP) in Subsection 3.1.7.

In a first step, the selection of vehicle functions was based on the requirement of a comparable degree of information content in the available instruction manuals. For the derivation of syntactic paraphrases, the extracted contents furthermore had to satisfy the requirement of semantic comparability. Based on these steps, four vehicle functions were selected (s. Table 3.3). Each of them is similar in its functional scope, in that it warns a driver and brakes the vehicle as soon as a certain condition is given. The sections in the instruction manual concerning the selected vehicle functions were compared in a next step in order to create a solid basis for structuring their content. Thereby, the similar structure of the content for the individual vehicle functions enabled the definition of instruction components that are necessary for explaining an assistant and its functionality. An overview of the result of this analysis is visualized in Figure 3.1. Here,

Table 3.3: Selected vehicle functions for the domain DAS. Adapted from Stier and Sigloch (2019, Table 1) with kind permission from Association for Computing Machinery.

1.	*Abstands-Assistent* ("Space Assist")
2.	*Nothalt-Assistent* ("Emergency Stop Assist")
3.	*Spurhalte-Assistent* ("Lane Keeping Assist")
4.	*Totwinkel-Assistent* ("Blind Spot Assist")

[1] Sources: German S-Class Owner's Manual (only available in the vehicle/at the dealer), Interactive Owner's Manual (`https://moba.i.daimler.com/baixn/cars/222.0_comand_2017/en_DE/index.html`; online: 09/02/2021), and an internal tool available to dealers for customer advice.

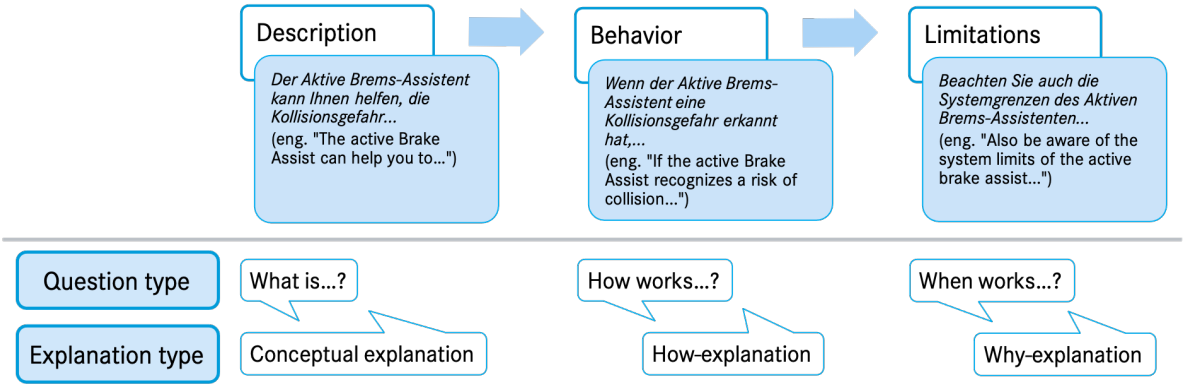


Figure 3.1: Components of an instruction manual and their mapping to question and answer types.

the first section of an instruction manual serves as an introduction and general description by briefly and concisely outlining a vehicle function's task and goal. This enables the discovery of a new concept and provides an answer to the question "What is function F?" This initial section is followed by an explanation of how the respective function works and behaves. This is done by means of describing a causal chain that causes a function to be performed, thus providing an answer to the question "How does function F work?" Finally, (external) influences are defined that lead to the deviation of the usual response behavior of the vehicle function. Various situations of constraints are listed here in response to the question "When can function F be used?" as a special case of the Why-question.

As mentioned before, the following preparation of syntactic paraphrases will focus on conceptual explanations for DAS functions in response to the question What. For this purpose, only the contents of the instruction manual component Description (s. Figure 3.1) are considered according to Grice's (1975) criterion of quantity.

Although the selected vehicle functions are similar in their functional scope, they differ in their orientation. As can be inferred from their names (s. Table 3.3), for example, the Space Assist aims to maintain a certain distance to the vehicle in front, while the Lane Keeping Assist keeps a vehicle in its lane. In order to nevertheless be able to create comparable explanations for the different vehicle functions in terms of their information density, it was therefore necessary to define a common information structure in a next step. Following the idea of functional design by Muthig and Schäflein-Armbruster (2008) for standardizing and

Table 3.4: Definition of a semantic-syntactic framework for the conceptual explanation following the model of FrameNet based on Baker *et al.* (1998) and its frame “Assistance”.

Frame: Conceptual explanation	
Description	A $benefited\_party$ benefits from a $tool$, which enables the $benefited\_party$ to achieve an abstract $purpose$. A $cause$ triggers the activation of the $tool$. The activation consists of one or more $action$s with the goal of achieving the concrete $purpose$ of the $tool$.
$Tool$	The $tool$ performs certain $action$s to benefit the $benefited\_party$.
$Benefited\_party$	The $benefited\_party$ is benefited by the $action$s of the $tool$.
$Cause$	The $cause$ identifies the actions of the $benefited\_party$ as the trigger for activating the $tool$ to perform an $action$.
$Action$	The $action$ by the $tool$ is triggered by the $cause$ to achieve the specific goal of the $tool$.
$Purpose$	The $purpose$ is an abstract goal of a desired state of the $benefited\_party$. The $purpose$ is the concrete goal of the $tool$ and describes the purpose of its functionality.

structuring information within text, the consistent design of explanations in the present context can be achieved by defining discourse elements as part of a uniform pattern. In addition to the structuring of content, its syntactic realization represents an essential part in this work. For this reason, a semantic-syntactic approach on the model of FrameNet (Baker *et al.*, 1998) was chosen, which will be explained in the following section.

3.1.4.2 Semantic-Syntactic Equivalence using FrameNet and the Criterion of Relation

With the help of the concept of a frame according to FrameNet (Baker *et al.*, 1998), the semantic-syntactic framework of a conceptual explanation was analyzed in more detail. The idea here was to define a semantic frame and to provide a syntactic framework of how individual frame elements interrelate. Thereby, a general structure of conceptual explanations was derived, which provided the basis for semantically and syntactically comparable explanations for different vehicle functions.

In contrast to FrameNet, however, this work is not intended to describe individual lexical units, but rather to define a frame for the construct of conceptual explanations. This frame can then provide information about which semantic content is assigned to the explanation and which syntactic roles can be taken within the frame. For this purpose, the instruction manual components of the individual vehicle functions selected in the previous section were compared in order to highlight similarities and differences between their functional elements. Subsequently, the contents were abstracted into the form of a frame following the example of FrameNet's frame "Assistance"[2] and semantic and syntactic roles were assigned to the individual frame elements (s. Table 3.4). Accordingly, the frame conceptual explanation consists of the five frame elements $Tool$ (T), $Benefited\_Party$ (B), $Cause$ (C), $Action$ (A_1, A_2) and $Purpose$ (P), which define its semantic-syntactic frame by their roles. Based on this frame, a general structure for conceptual explanations[3] was derived:

[Die Funktion F]$_T$ [kann [Sie]$_B$ [in einer bestimmten Situation]$_C$ unterstützen]$_{A_1}$.
[Die Funktion F]$_T$ [reagiert]$_{A_2}$, [um einem bestimmten Zweck zu dienen]$_P$.

As an answer to a question of the form "What is function F?" the conceptual explanation thus defines in which way a vehicle function provides support to the driver. The applied information structure (theme: function F, followed by rheme: definition of function F) ensures a comprehensible arrangement of the explanatory contents in the sense of the criterion of relation according to Grice (1975).

The deduced general structure ensures a semantic-syntactic equivalence when applied to prepare conceptual explanations in terms of different vehicle functions. Based on the developed frame, the instruction manual texts of the individual vehicle functions were arranged according to the frame elements. An example for a thereby resulting base text is provided in Example 3.3.[4]

[2] Source: `https://framenet.icsi.berkeley.edu/fndrupal/frameIndex` (Online: 10/02/2021)

[3] English translation:
[Function F]$_T$ [can support]$_{A_1}$ [you]$_B$ [in a particular situation]$_C$.
[Function F]$_T$ [reacts]$_{A_2}$ [to serve a specific purpose]$_P$.

[4] English translation:
[The active Brake Assist]$_T$ [can warn]$_{A_1}$ [you]$_B$ [at intersections and the ends of traffic jams]$_C$.
[It]$_T$ [brakes your vehicle if necessary]$_{A_2}$ [to avoid accidents with vehicles as well as pedestrians]$_P$.

Example 3.3 [Der aktive Brems-Assistent]$_T$ [kann [Sie]$_B$ [an Kreuzungen und Stauenden]$_C$ warnen]$_{A_1}$.
[Er]$_T$ [bremst Ihr Fahrzeug gegebenenfalls ab]$_{A_2}$, [um Unfälle mit Fahrzeugen sowie Fußgängern zu vermeiden]$_P$.

At this point, it should be emphasized that the frame conceptual explanation was specifically developed for the context of a definitional description. In general, this approach can also be applied to other types of explanations, for example, on the basis of the two instruction manual components behavior and limitations (s. Figure 3.1). However, due to the different focus of these explanations, a direct application of the frame conceptual explanation does not seem appropriate. For this purpose, separate frames would have to be created.

Despite the comparable semantic-syntactic base form of conceptual explanations, the individual frame elements differ across the vehicle functions, as they comprise different functionalities. To further ensure their comparable level of complexity at this point in terms of content and information density, several surface measures were employed in the following section.

3.1.4.3 Comparability in Content and Information Density and the Criterion of Manner

Following the criterion of manner according to Grice (1975), in a next step the comparability in terms of semantic complexity was ensured. For this purpose, different measures were applied on the created base texts of the individual vehicle functions.

Surface measures. On the sentence and word levels, the number of lexical units, their average number of characters, and the proportion of long words with more than 6 letters are considered reliable, easily interpretable measures. The syntactic complexity of a sentence increases with an increasing number of words. Similarly, lexical complexity increases with an increasing word length. The above measures were calculated for the base texts to ensure a comparable design in terms of their lexical and syntactic complexity.

Flesch Readability Index. The Flesch readability index according to Rudolf Flesch[5] is considered a qualitative measure for assessing how accessible a text is. Similar to the surface measures above, it assumes that the comprehensibility of a sentence increases with a decreasing word and sentence length. Based on Equation 3.1, the readability index for German

[5] `https://fleschindex.de/formel/` (Online: 10/02/2021)

is computed from the average sentence length as the number of words per sentence and the average number of syllables per word. The higher the calculated readability ease score is within a range from 0 to 100, the more comprehensible a text is considered to be.

$$Flesch\ index = 180 - Av.\ sentence\ length - (58.5 * Av.\ number\ of\ syllables) \tag{3.1}$$

Information Density. In order to account for the different functionalities of different vehicle functions and thereby varying lexical units within the individual frame elements, the information or idea density approach was employed. The idea density is based on the insight that texts with low information density are easier to understand (Kintsch and Keenan, 1973), since each idea or proposition in a text requires a certain amount of processing effort (Covington, 2009). In the context of preparing comparable conceptual explanations, the propositional analysis represents a valuable method to measure complexity in terms of information density and to ensure comparability across different vehicle functions by means of a similar number of propositions. In general, there are various approaches to defining a proposition and calculating information density (*e.g.*, Kintsch and Keenan, 1973; Turner and Greene, 1977). Since the semantic complexity of a conceptual explanation may depend on single lexical units, this work relies on a phase-based approach upon surface structure relations. The propositional analysis was performed on the basis of the guideline by Chand *et al.* (2012), who consider the propositions of a text by means of a dependency-like structure. Following the authors, information density was calculated as the number of propositions per ten words (s. Equation 3.2).

$$Information\ density = \left(\frac{Number\ of\ propositions}{Number\ of\ words} \right) * 10 \tag{3.2}$$

The measures indicated above were computed for the base texts of all vehicle functions (s. Table 3.5). Due to the careful execution from the content selection to its semantic-syntactic structuring, the calculated values revealed a comparability of the base texts in terms of their lexical and semantic complexity. The absolute word count ranges between 23 and 25 words ($M = 23.75$, $SD = 0.83$) with an average word length of 5.88 characters ($SD = 0.07$). The proportion of long words $>$6 characters is found between 26-35% ($M = 0.31$, $SD = 0.04$). On average, between three and four propositions are used, or ideas introduced, per 10 words in the base texts ($M = 3.79$, $SD = 0.27$). An interpretation of these values can be based on the Flesch index ($M = 53$, $SD = 1.22$). According to this readability score, the created base texts are thus considered to be challenging. Overall, the calculated values generally

Table 3.5: Overview of qualitative measures computed for the DAS vehicle functions.

	Base text / MCV	**WC**	**WL**	**BW**	**ID**	**FI**
Abstands-Assistent (Space Assist)	*Der aktive Abstands-Assistent kann Sie bei zu dichtem Auffahren warnen. Er bremst Ihr Fahrzeug gegebenenfalls ab, um den Abstand zu vorausfahrenden Fahrzeugen zu regeln.* (eng. "The active Space Assistant can warn you if you are driving too close. If necessary, it brakes your vehicle to regulate the distance to vehicles in front.")	24	5.92	0.33	4.17	53
Nothalt-Assistent (Emergency Stop Assist)	*Der Aktive Nothalt-Assistent kann Sie bei dauerhafter Ablenkung warnen. Er bremst Ihr Fahrzeug kontrolliert bis zum Stillstand ab, um eine Kollision zu verhindern.* (eng. "The active Emergency Stop Assist can warn you if you are permanently distracted. It brakes your vehicle to a standstill in a controlled manner to prevent a collision.")	23	5.96	0.35	3.48	54
Spurhalte-Assistent (Lane Keeping Assist)	*Der aktive Spurhalte-Assistent kann Sie bei unbeabsichtigtem Verlassen der Fahrspur warnen. Er bremst eigenständig, um Ihr Fahrzeug zurück in die Spur zu führen.* (eng. "The active Lane Keeping Assist can warn you if you leave your lane unintentionally. It brakes independently to guide your vehicle back into the lane.")	23	5.87	0.26	3.91	54
Totwinkel-Assistent (Blind Spot Assist)	*Der aktive Totwinkel-Assistent kann Sie bei einem Spurwechsel vor Fahrzeugen im toten Winkel warnen. Er bremst Ihr Fahrzeug eigenständig ab, um eine Kollision zu vermeiden.* (eng. "The active Blind Spot Assist can warn you of vehicles in your blind spot when you change lanes. It brakes your vehicle independently to avoid a collision.")	25	5.76	0.28	3.60	51
	M (SD)	23.75 (0.83)	5.88 (0.07)	0.31 (0.04)	3.79 (0.27)	53 (1.22)

Note: WC- word count, WL- av. word length (in characters), BW- proportion of big words ($>$ 6 characters), ID- idea density, FI- flesch index

vary ± 1 SD from the mean, generally indicating a consistent behavior across all vehicle functions. However, in the case of conspicuously deviating values, minor manual modifications were intended. In particular, the idea density of 4.17 for the Space Assist was notably high compared to the other vehicle functions ($M = 3.79$, $SD = 0.27$). The reason for this can be found, for example, in the number of propositions within the frame element $Purpose$. While the Space Assist has the task to "regulate the distance to vehicles in front" (3 propositions), the Emergency Stop and Blind Spot Assists have the clear goal to "avoid/prevent a collision" (1 proposition). The more detailed definition of the specific purpose of the Space Assist is realized here by an additional prepositional phrase (*zu vorausfahrenden Fahrzeugen*; eng. "to vehicles in front"), which is also found, for example, in the Lane Keeping Assist (*zurück in die Spur*; eng. "back into the lane"). Omitting these phrases would reduce the information density of the frame element $Purpose$ and thus of the two functions Space and Lane Keeping Assist and align them with the level of Emergency Stop and Blind Spot Assist. However, this would contradict the criterion of quantity according to Grice (1975). For this reason, no manual adjustment of the created base texts was made.

Rather, the texts created up to this point for the four different DAS functions represent a basis for subsequent syntactic paraphrases. By the careful execution of the creation process of these base texts, they fulfill all previously defined requirements, including Grice's criteria (1975), for conceptual explanations in the context of one-shot QAS in the vehicle. As shown in Table 3.5, the conceptual explanations of the driving assistants follow the pattern of a definitional explanation in the form of two concise sentences. Due to their comparable semantic-syntactic structure based on the specifically created frame for a conceptual explanation and their comparability in terms of semantic complexity, the basic texts presented above serve as a starting pattern for further syntactic realizations and are therefore referred to below as the first syntactic variant, the Main Clause Variant.

3.1.4.4 Syntactic Variation through Aggregation

The above process was used to ensure the comparability of the information structure as well as a comparable information density of the individual DAS explanations. In the following, the above defined base texts in the form of a Main Clause Variant (MCV) were paraphrased to generate syntactic variants.

For the purpose of syntactic paraphrasing, several aggregation strategies were applied.

Table 3.6: Overview of syntactic paraphrase variants.

1.	Main clause variant	MCV
2.	Final clause variant	FCV
3.	Nominal clause variant	NCV
4.	Relative clause variant	RCV

Given the complexity of language (*e.g.*, Winograd, 1972; Moore, 2017), the selection in this work is by no means to be considered exhaustive. Rather, the selection was based on the fact that different syntactic forms exhibit different degrees of syntactic complexity (s. Section 2.2). Therefore, in order to be able to investigate the influence of syntactic realizations in voice output on the perceived naturalness and comprehensibility, aggregation strategies were selected to paraphrase conceptual explanations with different levels of syntactic complexity. The selection of aggregation mechanisms was inspired by previous work in this area, notably by Florencio *et al.* (2008) and Mairesse and Walker (2011), whose use of different aggregation strategies was grounded in more natural output in NLG systems.

In this work, syntactic aggregation is understood as the process of combining constituents by means of syntactic rules, such as through coordination or subordination (Florencio *et al.*, 2008). By means of these aggregation strategies, four syntactic paraphrases (s. Table 3.6) were created on the basis of the MCV base texts by restructuring frame elements (s. Table 3.2 and the derived general structure). Lexical material was kept constant and ungrammatical paraphrases were excluded. The structure of the resulting paraphrases is outlined below.

MCV: [T A_1 [B C]] [T A_2 P]
FCV: [T A_1 [B C] t A_2 P]
NCV: [T A_1 [B C] t P [$A_{2,\ \text{nominalized}}$]]]
RCV: [T [t B C A_1] A_2 P]

The final clause variant (FCV) is characterized by the coordination of the two MCV clauses and an internal subject ellipsis. The nominal clause variant (NCV) consists of the same coordination strategy with subject ellipsis. In addition, it includes an infinitive nominalization as desententialized connective means within the frame element $Action_2$. The relative clause variant (RCV) includes a subject-oriented subordination of the frame elements $Action_1$,

Benefited_party and *Cause* with a corresponding restructuring of the frame elements to a verb-to-last position. The syntactic paraphrasing according to the patterns described above was performed equally for the individual vehicle functions of the domain DAS. By following this procedure, the semantic-syntactic comparability of the resulting explanations should be maintained and, in addition, the syntactic complexity should be prevented from changing between the individual driving assistants. The resulting syntactic paraphrases are exemplarily demonstrated in Table 3.7. An overview of the syntactic paraphrases for the different vehicle functions is found in Appendix A.1.

Due to the aplication of different aggregation strategies and paraphrasing of individual constituents, an increasing complexity of the syntactic paraphrases is assumed. They thus form a continuum of syntactic complexity from a simple, linear structure in MCV to a nested syntactic structure in RCV:

MCV	$<$	**FCV**	$<$	**NCV**	$<$	**RCV**
parataxis, linear structure		coordination, inner subject ellipsis		coordination, inner subject ellipsis, infinitive nominalization		subject-oriented relative clause

3.1.5 Pilot Study 1: Validating the Approach to Prepare Syntactic Paraphrases

A pilot study was conducted in order to validate the generation approach and the resulting syntactic paraphrases. The focus here was particularly on whether the created paraphrases fulfilled the criteria defined by Grice (1975) (s. Subsection 3.1.3). The following subsection is based on the publication by Stier and Sigloch (2019) and provides a description of the employed methodology, before the results are presented.

3.1.5.1 Methodology and Experimental Design

All participants were Daimler AG employees with German as their native language and voluntarily participated in this pilot study. Before starting the study, they were asked to provide demographic data, such as age and gender. Furthermore, they were introduced into the study procedure, without referring to the aspect of varying syntactic forms.

Table 3.7: Realization of different aggregation strategies, demonstrated for the Brake Assist.

MCV (base)	*Der aktive Brems-Assistent kann Sie an Kreuzungen und Stauenden warnen. Er bremst Ihr Fahrzeug gegebenenfalls ab, um Unfälle mit Fahrzeugen sowie Fußgängern zu vermeiden.* (eng. "The active Brake Assist can warn you at intersections and the ends of traffic jams. It brakes your vehicle if necessary to avoid accidents with vehicles as well as pedestrians.")
FCV	*Der aktive Brems-Assistent kann Sie an Kreuzungen und Stauenden warnen und bremst Ihr Fahrzeug gegebenenfalls ab, um Unfälle mit Fahrzeugen sowie Fußgängern zu vermeiden.* (eng. "...and brakes your vehicle if necessary to avoid accidents...")
NCV	*Der aktive Brems-Assistent kann Sie an Kreuzungen und Stauenden warnen und vermeidet durch Abbremsen Ihres Fahrzeugs Unfälle mit Fahrzeugen sowie Fußgängern.* (eng. "...and avoids accidents ... by braking your vehicle.")
RCV	*Der aktive Brems-Assistent, der Sie an Kreuzungen und Stauenden warnen kann, bremst Ihr Fahrzeug gegebenenfalls ab, um Unfälle mit Fahrzeugen sowie Fußgängern zu vermeiden.* (eng. "The active Brake Assist, which can warn you at intersections and the ends of traffic jams, brakes your vehicle if necessary to avoid accidents with vehicles as well as pedestrians.")

This pilot study was conducted without a parallel secondary task to ensure that the participants were able to concentrate solely on their primary task, that is, listening to and evaluating the manually created paraphrases. It took place in the form of a semi-guided interview (s. Appendix A.2.1) in an examination room, with the particpants sitting at a table opposite their investigator. In the context of one-shot QAS, the participants were asked to subsequently formulate questions in the form "What is...?" concerning the four DAS functions to an imaginary SDS. As a system response, the synthesized conceptual explanations were played by the investigator and the participants were asked to evaluate them on a 3-point Likert scale. The experimenter was responsible for documenting their answers and comments.

As depicted in Figure 3.2, in a first step, one syntactic realization for the four vehicle functions was played back to the participants. Accordingly, the participants were asked to rate the completeness and appropriateness of the content for each explanation in turn, and then the uniformity of structure and form across the functions. This procedure was randomly repeated for the different syntactic realizations of the conceptual explanations. In a next step, the par-

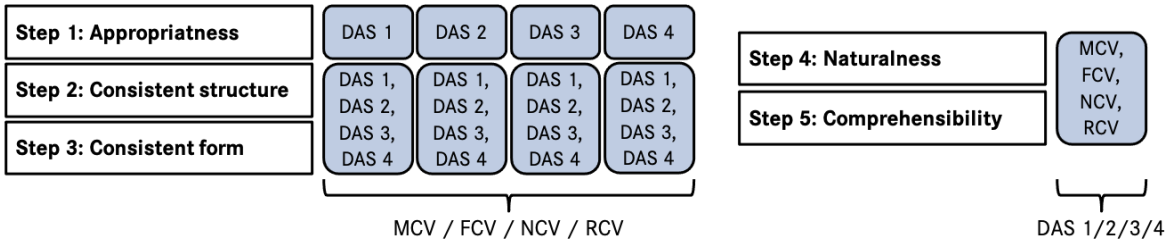


Figure 3.2: Schematic procedure of Pilot Study 1.

ticipants were presented with the different syntactic realizations of an explanation. They were asked to rate each paraphrase in terms of its perceived comprehensibility and naturalness and indicate their preferred variant. This procedure was repeated for the randomized DAS functions.

3.1.5.2 Results

A total of 18 German native speakers between 23 and 63 years ($M = 30.0$, $SD = 9.59$) participated in the experiment. They comprised 11 male and 7 female subjects.

All participants found the content provided in the voice prompts to be appropriate and sufficiently comprehensive for the context of conceptual explanations. Furthermore, they recognized a consistent structure and phrasing of the content between the explanations of the individual DAS functions. In addition, a clear differentiation between and prioritization of syntactic paraphrases emerged through the investigation concerning the perceived naturalness and comprehensibility of the explanations. As summarized in Figure 3.3, the participants indicated a high acceptance of MCV (26%) compared to FCV (16%), NCV (12%) and RCV (8%).

The observations described above allowed to answer the research questions set in this pilot study. Overall, a consistent structure of the explanations for the different vehicle functions was recognized by all participants, thus confirming their syntactic-semantic comparability. The validity of the approach to produce syntactic paraphrases was thus validated. In addition, the ability of the participants to distinguish and prioritize different syntactic forms with respect to their stated preferences was observed. The extent to which this differentiation takes place

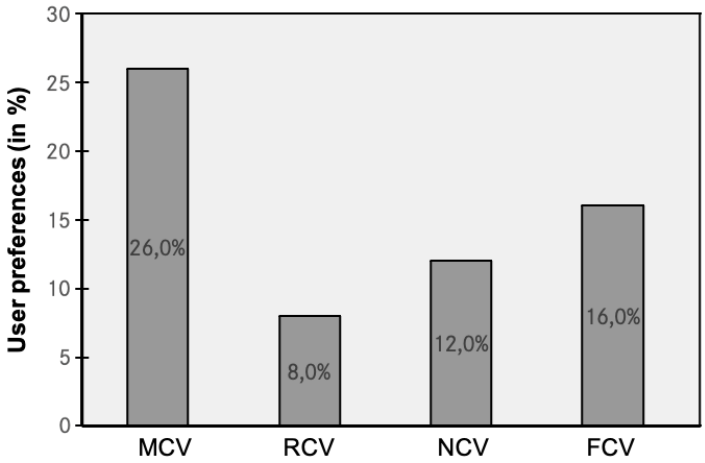


Figure 3.3: Indicated user preferences of different syntactic forms in Pilot Study 1. Adapted from Stier and Sigloch (2019, Figure 1) with kind permission from Association for Computing Machinery.

(un)consciously by listening to voice output was investigated in a further pilot study.

3.1.6 Pilot Study 2: Investigating the Level of Consciousness

It is common practice to assess preferences for syntactic structures by means of text samples (*e.g.*, Mairesse and Walker, 2011). However, in the context of evaluating voice output, the audio channel represents a decisive factor. Particularly in the case of in-vehicle SDS prompts, it is expected that their applicability needs to be evaluated in the interaction context of driving and thus be delivered to the driver via the audio channel in a lifelike manner. To the best of our knowledge, however, there is no prior work nor evidence on whether prompts received via audio can be used to collect valid user preferences with respect to their syntactic form. For this reason, a second pilot study was conducted to investigate whether participants are able to intuitively distinguish or even explicitly identify differences between varying syntactic realizations in voice output, that is, via audio. It is hypothesized in this context that user preferences are influenced by an awareness for syntactic forms due to established opinions concerning their appropriateness. This pilot study thus examines whether this (lack of) awareness allows intuitve and unbiased user ratings.

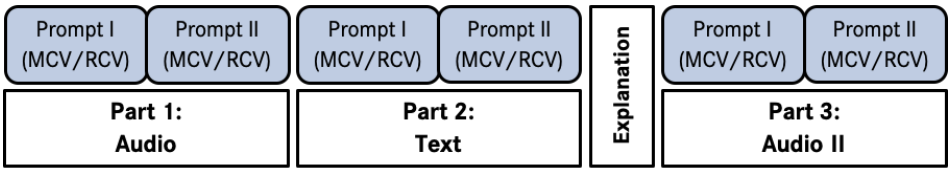


Figure 3.4: Schematic procedure of Pilot Study 2. Taken from Stier *et al.* (2020b, Figure 2).

The work presented in this section is based on the publication by Stier *et al.* (2020b). In the following, a desciption of the employed methodology and experimental design are provided, before the results of this pilot study are presented.

3.1.6.1 Methodology and Experimental Design

Prior to the study, the participants were asked to provide demographic information, such as age and gender. Subsequently, they were introduced to follow the instructions of the experimenter. No further preparation took place.

In order to focus the participants' attention on the syntactic differences, the pilot study was conducted without a parallel secondary task. It took place in the form of a semi-guided interview (s. Appendix A.2.2) in an examination room, with the participants sitting at a table opposite their investigator. The experimenter was responsible for guiding the participants through the procedure visualized in Figure 3.4 and documenting their answers: In Part 1, the participants were asked to listen to one MCV and one RCV as synthesized voice prompts in random order for two different DAS functions. The experimenter then subsequently presented prompts in text form for two further DAS functions to the participants in the identical order of the syntactic forms (Part 2) and afterwards revealed the differences between the two syntactic variants in the Explanation phase. In Part 3, the particpants were then asked to listen to the synthesized voice prompts again. After each voice or text prompt they were asked whether they noticed any peculiarity. Additionally, after the completion of each part, they were asked which of the two variants they preferred.

The order of the syntactic features was randomized over participants. In contrast, the order of the overal study procedure presented here was deliberately not randomized in order to investigate the perception of syntactic forms and to observe to what extent user preferences

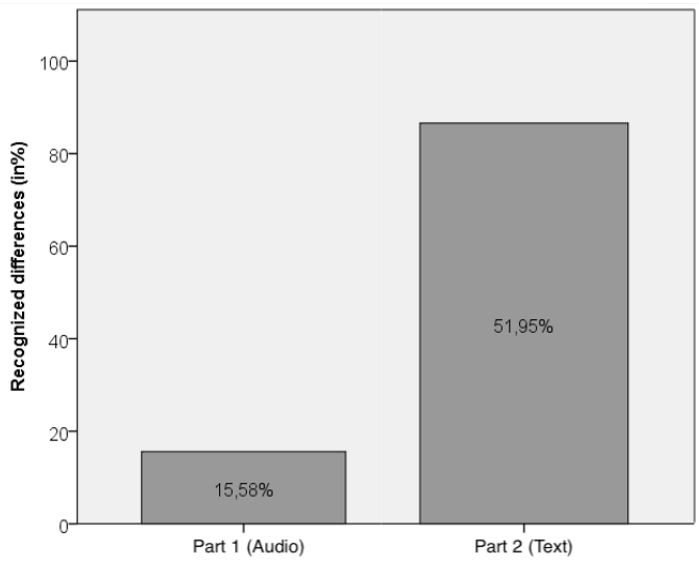


Figure 3.5: Recognized differences between syntactic paraphrases.

change depending on the awareness of syntactic structures. Randomization of the individual parts would prevent these observations.

The syntactic paraphrases employed in this pilot study either as voice prompt or in text form are provided in Appendix A.2.3 and A.2.4, respectively.

3.1.6.2 Results

Overall, 77 German native speakers between 18 and 69 years ($M = 43.60$, $Mdn = 45$, $SD = 14.65$) participated in this pilot study. They comprised 46 male and 31 female subjects.

Figure 3.5 provides an overview of the recognized differences between the tested paraphrases. While only 12 participants (15.58%) explicitly identified the syntactic differences via audio in Part 1, a clear majority of 40 participants (51.95%) perceived them through text in Part 2. On a conscious level, the syntactic differences were therefore perceived significantly less as a distinguishing feature via the audio channel than in the text (Wilcoxon signed-rank test, $Z = -5.292$, $p < .001$, $r = .60$).

Furthermore, a shift in the evaluation behavior of the participants throughout the pilot study

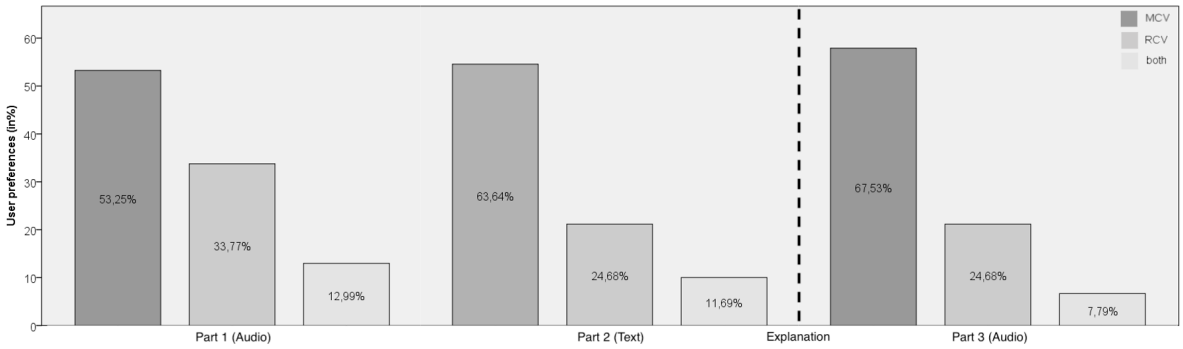


Figure 3.6: Indicated user preferences. Adapted from Stier *et al.* (2020b, Figure 3), © 2021 Copyright held by the owner/author(s).

became apparent (Figure 3.6). While no clear difference in the assessed user preferences was revealed between Part 1 and Part 2 ($Z =$ -1.949, $p = .051$, $r = .16$), they differed significantly between Parts 1 and 3, that is, *before* and *after* the Explanation phase ($Z =$ -3.192, $p < .001$, $r = .26$). The awareness for the syntactic differences after the explanation phase led to an increased preference for MCV (Part 1: 53.25%, Part 3: 67.53%) and a decreased preference for RCV (Part 1: 33.77%, Part 3: 24.68%). When questioned about their changed evaluation behavior in this context, the participants stated that a relative clause would be too complex and therefore rather unsuitable in voice output.

The observations above generally confirm the research hypothesis underlying this pilot study that the awareness for syntactic forms does influence user preferences. It became apparent that the perception of syntactic differences over audio was lacking for a large majority of participants via audio in Part 1. However, at the same time, participants subconsciously prioritized the presented syntactic structures and indicated clear preferences. Apparently, it is precisely this lack of awareness which seems to allow for intuitive user ratings without a fixed opinion on syntactic complexity and its applicability. From the results of this pilot study, it is thus concluded that the assessment of voice output via audio concerning individual syntactic preferences represents a valid methodology in our context.

3.1.7 Application and Extension of the Approach

In the described approach, syntactic paraphrases for conceptual explanations were created on the basis of the domain DAS under consideration of Grice's criteria (1975). It is based on the definition of an information structure and the semantic-syntactic comparability of the individual explanations. In addition, semantic complexity was ensured by means of objective measures. In order to investigate the influence of syntactic forms in different domains in the following, the approach was applied to another domain and thus extended. For this purpose, comfort programs (COP) were chosen as a second domain from the application context entertainment in contrast to the domain DAS. Whereas driving assistants provide security-related aids while driving, the comfort functions focus on the well-being and efficiency of a driver and include a composition of, for example, fragrance, lighting and massage functions.

The syntactic paraphrases for the domain COP were created analogously to those of the domain DAS under consideration of the requirements defined in Sections 3.1.1–3.1.3. In a first step, four vehicle functions with a similar functional scope of providing an interplay of supportive programs were selected (s. Table 3.8). Their contents were assigned to different instruction manual components (s. Subsection 3.1.4.1) and only the information providing a definitional conceptual explanation to the question What was considered further. In a next step, the COP content was mapped to the frame conceptual explanation (s. Subsection 3.1.4.2). Besides different verb phrases ($Action_1$: support vs. serve; $Action_2$: react vs. do) and a corresponding customization of the realized objects due to the different functional backgrounds of the DAS and COP functions, the semantic-syntactic structure could generally be adopted for conceptual explanations of both domains:[6]

Table 3.8: Selected vehicle functions for the domain COP. Adapted from Stier and Sigloch (2019, Table 1) with kind permission from Association for Computing Machinery.

1.	*Behaglichkeit* ("Well-being")
2.	*Freude* ("Joy")
3.	*Vitalität* ("Vitality")
4.	*Wärme* ("Warmth")

[6] English translation:
[Function F]$_T$ [can serve]$_{A_1}$ [you]$_B$ [in a particular situation]$_C$.
[Function F]$_T$ [does something]$_{A_2}$ [to create a particular state]$_P$.

DAS: [Die Funktion F]$_T$ [kann [Sie]$_B$ [in einer bestimmten Situation]$_C$ unterstützen]$_{A_1}$.
[Die Funktion F]$_T$ [reagiert]$_{A_2}$, [um einem bestimmten Zweck zu dienen]$_P$.

COP: [Die Funktion F]$_T$ [kann [Ihnen]$_B$ [in einer bestimmten Situation]$_C$ dienen]$_{A_1}$.
[Die Funktion F]$_T$ [tut etwas]$_{A_2}$, [um einen Zustand zu erzeugen]$_P$.

A similar semantic complexity was therefore expected between the individual COP functions and the domains DAS and COP in general by means of objective measures (s. Subsection 3.1.4.3). An overview of the computed measures for the COP base texts is provided in Table 3.9. Similar to DAS, the base text lengths for COP range between 22 and 25 words ($M = 23.50$, $SD = 1.12$) with an average word length of 6.08 characters ($SD = 0.31$) and a proportion of big words between 36-50% ($>$6 characters; $M = 0.42$, $SD = 0.05$). On average, four propositions are introduced per 10 words within a range from 3.75 to 4.35 ($M = 4.05$, $SD = 0.21$). According to the Flesch index ($M = 37.5$, $SD = 3.77$), the COP base texts are considered to be difficult. Overall, also for the domain COP, a difference between individual functions was evident concerning their idea density values. This observation can for instance again be attributed to the number of propositions within the frame element $Purpose$, here due to the additional use of an adjective. While the function Joy with an idea density value of 4.35 aims to “create a positive mood” (2 propositions), the program Well-being tries to “enhance your well-being” (1 proposition) with an information density of 3.75. Similarly complex noun phrases are also found in the functions Vitality and Warmth (“invigorating effect”, “comfortable ambience”). As for DAS, a manual modification of phrases was omitted since the COP texts generally indicate homogeneous values in the computed measures and in order to ensure the criterion of quantity according to Grice (1975). Analogous to the domain DAS, as described in Subsection 3.1.4.4, syntactic paraphrases were constructed on the basis of the COP base texts (s. Appendix A.2).

Overall, the base texts of the DAS and COP domains differ only slightly in their average length. Besides a greater variance in text lengths for COP, its functions are mainly characterized by the use of generally longer words. Although the COP texts appeared more consistent overall than those of the domain DAS concernig their complexity in terms of information density, they are considered to be more complex from a holicstic perspective taking their idea density and Flesch index into account. Due to the identical preparation of the base texts for the two domains COP and DAS, this different degree of semantic complexity is attributed to the different nature of their textual content and functioning. While the COP texts particularly use adjectives that trigger emotions (*e.g.*, tense, stressful, relaxing, comfortable) to describe rather

Table 3.9: Overview of qualitative measures computed for the COP vehicle functions.

	Base text / MCV	WC	WL	BW	ID	FI
Behaglichkeit (Well-being)	*Das Programm Behaglichkeit kann Ihrem Komfort in angespannten Fahrsituationen dienen. Es erzeugt durch eine wärmende Massage ein entspanntes Spa-Feeling, um Ihr Wohlbefinden zu steigern.* (eng. "The program Well-being can serve your comfort in tense driving situations. It creates a relaxing spa feeling through a warming massage to enhance your well-being.")	24	6.58	0.50	3.75	34
Freude (Joy)	*Das Programm Freude kann Ihrem Komfort in ermüdenden Fahrsituationen dienen. Es nutzt eine aktivierende Sitzmassage und Musik, um eine positive Stimmung zu erzeugen.* (eng. "The program Joy can serve your comfort in tiring driving situations. It uses an activating seat massage and music to create a positive mood.")	23	6.09	0.39	4.35	34
Vitalität (Vitality)	*Das Programm Vitalität kann Ihrem Komfort in monotonen Fahrsituationen dienen. Es nutzt anregendes Licht und Musik, um eine belebende Wirkung zu erzeugen.* (eng. "The program Vitality can serve your comfort in monotonous driving situations. It uses stimulating light and music to create an invigorating effect.")	22	5.91	0.41	4.09	39
Wärme (Warmth)	*Das Programm Wärme kann Ihrem Komfort in belastenden Fahrsituationen dienen. Es erzeugt durch beheizte Sitze eine wohlige Wärme, um für ein gemütliches Ambiente zu sorgen.* (eng. "The program Warmth can serve your comfort in stressful driving situations. It creates a cozy warmth through heated seats to provide a comfortable ambience.")	25	5.76	0.36	4.00	43
	M (SD)	23.50 (1.12)	6.08 (0.31)	0.42 (0.05)	4.05 (0.21)	37.5 (3.77)

Note: WC- word count, WL- av. word length (in characters), BW- proportion of big words ($>$ 6 characters), ID- idea density, FI- flesch index

abstract concepts of the comfort functions (*e.g.*, invigorating effect, comfortable ambience), the DAS texts are limited to comparatively unemotional descriptions of how the driving assistants work. Although the approach of preparing conceptual explanations is applicable to different domains, it has been shown that an equivalence of the resulting base texts with respect to their semantic-syntactic complexity is only conditionally possible. The extent to which these divergent backgrounds of different domains affect the perception of syntactic paraphrases in the vehicle will be investigated in the following user studies.

In this section, the foundation for investigating the influence of syntactic forms in voice output was established. The approach to create syntactic paraphrases within the context of one-shot conceptual QAS based on their semantic-syntactic comparability was validated by means of a first pilot study. In a second pilot study, it was furthermore demonstrated that the lack of awareness of syntactic differences in voice prompts allows the elicitation of valid user preferences with respect to their syntactic form. Based on these findings, two driving simulator studies were subsequently conducted to investigate the influence of syntactic forms in vehicle-based voice output.

3.2 Investigating the Influence of Syntax in a Dual-Task Environment

On the basis of the work presented in Section 3.1, an exploratory Wizard-of-Oz (WoZ) experiment was conducted in a driving simulator. The goal of this study was to identify parameters, which may be related with the perception of in-vehicle prompts and to reveal the influence of syntactic forms in voice output. In this respect, this study is considered basic research in order to approach the long-term question of how system-side explanations as one example of linguistically complex voice output should be designed to be perceived as natural and intuitively comprehensible for an individual driver. For this purpose, the participants had to interact with a simulated SDS while driving and evaluate syntactically differing voice prompts embedded in one-shot QAS for the question type What with respect to their perceived naturalness and comprehensibility.

The following sections, which are based on the publication by Stier and Sigloch (2019), provide a detailed description of the employed methodology, before the results are presented and discussed.

3.2.1 Methodology

This section describes the methodological approach of the user study. First, the participants and experimental design are presented. Finally, the employed materials and the procedure are introduced.

3.2.1.1 Participants

A total of 48 native speakers of German participated in the experiment with an average age of 38.15 years ($SD = 14.17$) and a gender distribution of 36 male and 12 female subjects. All of them possessed a valid driver's license. The participants received an expense allowance of 50€ each for their participation.

3.2.1.2 Experimental Design

This simulator study had a deliberately exploratory character. It was therefore not designed to test concrete hypotheses, but to answer a series of openly formulated research questions:

RQ 1 Is the perception of voice output influenced by user and system parameters? In particular, does the syntactic form of voice prompts play a role in a driving context?

RQ 2 Which user- and system-sided parameters generally influence the perceived comprehensibility and naturalness of syntactically differing voice prompts?

In order to answer the questions above, a QAS for the question type What was chosen as interaction environment for this user study. Furthermore, various controllable system-related parameters were considered as possible influencing factors on the perception of syntactically different prompts. Figure 3.7 provides an overview of these selected parameters, which are additionally explained in the following paragraphs.

Sentence type. Syntactic forms differ in their inherent complexity (s. Section 2.2). Thus, one main objective of this study was to investigate whether there is a direct impact of syntactic forms and their inherent complexity on the subjective perception of a driver due to the different level of cognitive load they induce. For this purpose, explanatory voice prompts were included in this user study in a total of four different syntactic realizations with increasing complexity (MCV $<$ FCV $<$ NCV $<$ RCV). Their manual preparation is described in Section 3.1.

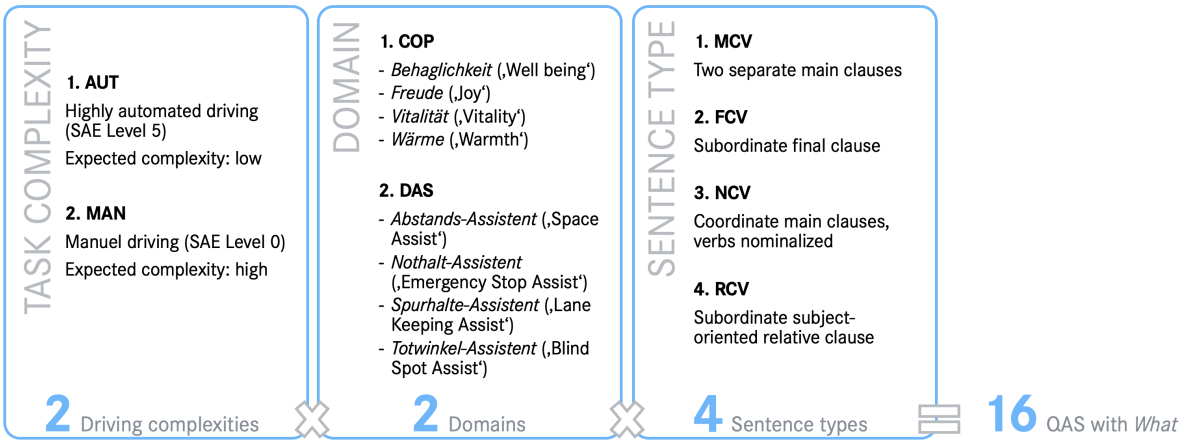


Figure 3.7: A set of parameters was considered in the experimental design. Overall, each participant experienced the QAS What 16 times.

Domain. The level of familiarity may influence whether a content is intuitively understood (s. Section 2.3). For this reason, the two different domains related to the vehicle and driving context COP and DAS were chosen as prompt contents (s. Section 3.1). DAS and COP differ in terms of their applicable contexts (security vs. entertainment), contents (driving aids vs. comfort functions) and degree of prominence (well-established, known vs. recently introduced commercially).[7] Furthermore, the functional scope of DAS can be derived directly from their name, whereas this is not the case for the individual functions of COP.

Syntactic paraphrases were realized for both domains DAS and COP, that is including four different in-vehicle functions per domain (s. Figure 3.7). Although the individual functions are part of this user study, they only serve to evaluate the different sentence types without creating any sequence effects. The realization of different functions furthermore served to provide content variability so that the participants' attention was not diminished by the task of repeatedly listening to the same contents. Therefore, the functions are listed here for completeness, but are not examined as a separate influencing factor in this study.

[7] The comfort programs were first launched on the market in September 2017, while the driving assistants have already had a long history in the vehicle, at the latest since the Brake Assist system was introduced in series production in 1996. Sources:
`https://media.daimler.com/marsMediaSite/de/instance/ko/ENERGIZING-Komfortsteuerung-Wellness-beim-Fahren.xhtml?oid=22934464` (Online: 01/12/2020)
`https://www.daimler.com/innovation/specials/chronologie-der-assistenzsysteme.html` (Online: 02/12/2020)

Driving complexity. The cognitive load of a driver decreases with the level of vehicle automation (s. Section 2.3). For this reason, the driving task in the driving simulator included both one manual (MAN) and one autonomous (AUT) part. These two opposing driving situations were selected on the basis of their differing complexity due to their diverging levels of cognitive demand: While the participant was supposed to take over the controls during manual driving (SAE Level 0, *cf.* SAE, 2018), including steering, braking and accelerating, the vehicle performed these tasks during autonomous driving (SAE Level 5).

The parameters described above were integrated as factors into the experimental within design in order to investigate their influence on the perception of syntactically differing voice prompts. As indicated in Figure 3.7, each participant experienced 16 QAS with the question type What in two driving complexity levels (AUT, MAN) x two domains (COP, DAS) x four sentence types (MCV, FCV, NCV, RCV). In addition to the controllable system parameters listed here, personal information of the participants was collected by means of questionnaires in order to investigate the influence of personal characteristics on the perception of voice output in the vehicle. They are described in the following section.

3.2.1.3 Materials

This subsection introduces the materials, which were employed in the user study.

3.2.1.3.1 Questionnaires. In this user study, several demographic and personal information about the participants was collected in order to identify the influence of these parameters on the perception of voice output. In the course of a preliminary survey, three different questionnaires were used. They were created using a text processing program and presented to each subject in printed form. The applied questionnaires can be found in Appendix A.3.1.

- **Preliminary Questionnaire:** Humans differ with respect to their individual attributes, such as age, gender, and prior experiences. Thus, in a first step the subjects were asked to provide their demographic data (age, gender, etc.). Additionally, they were asked to self-assess their level of linguistic knowledge and prior experiences with the domains DAS and COP on a 5-point Likert scale.
- **Technical Affinity:** Given the context of QAS concerning in-vehicle functions, the degree of technical affinity in general is assumed to represent a distinguishing factor. The

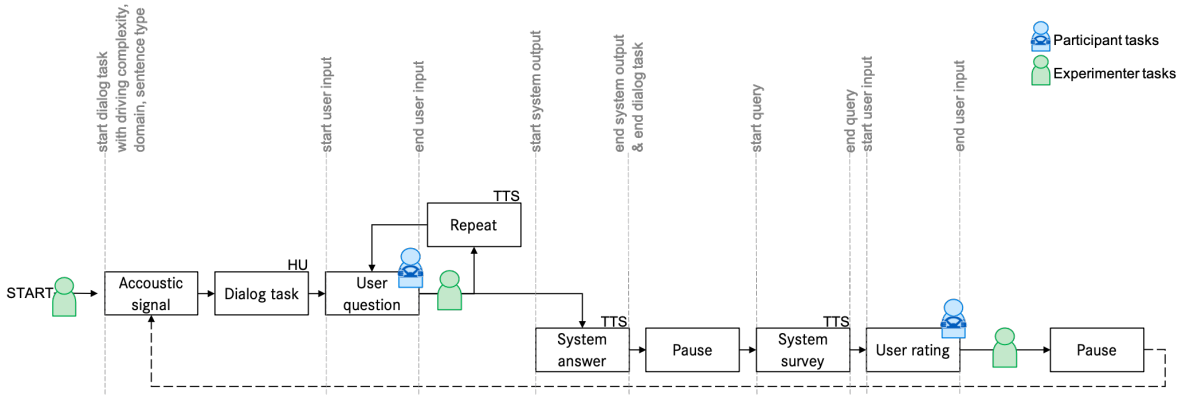


Figure 3.8: Schematic representation of a dialog flow using the WoZ tool.

questionnaire described by Karrer *et al.* (2009) served as a standardized procedure to capture the affinity for technology, which is defined as a characteristic of personality that expresses itself in a person's positive attitude, enthusiasm and trust towards technology (s. Section 2.4). The questionnaire consists of 19 items, each assigned to one of four components. The questions were asked on a 5-point scale.

– **Big Five Personality Traits:** The participants were additionally asked to self-assess their personality traits on a 5-point Likert scale using the German version of the Big Five Inventory (BFI; *e.g.*, Goldberg, 1990; s. Section 2.4) questionnaire validated by Rammstedt and Danner (2016). The questionnaire was used to be able to describe the wide range of human behavior in a few dimensions. It consists of 45 questions relating to a total of five personality traits.

3.2.1.3.2 Wizard-of-Oz as Simulated Spoken Dialog System. The basis for this WoZ experiment was a Daimler internal tool created on the model of SUEDE (Klemmer *et al.*, 2000) to simulate the spoken interaction between a user and a real SDS. For this purpose, the WoZ tool provides the functionality to specify the required dialog flow on the one hand and a working environment for the experimenter on the other.

The specification of the dialog flow in the WoZ tool is done by means of action states. They can be arranged and linked via a graphical user interface. In addition, each state can be assigned one or more actions, such as logging state information or realizing system reactions

like generating voice output. During the experiment, the WoZ tool served as an SDS in which the ASR and NLU were replaced by the experimenter (*i.e.*, the wizard). Depending on the dialog state and the participant's reaction, the wizard interacted with the WoZ tool via buttons to move to the next dialog state. Figure 3.8 shows an example of a dialog flow with the WoZ tool: After the experimenter (green) started the dialog via a corresponding button, an acoustic signal indicated that a new dialog task for the participant was displayed in the head-unit (HU) screen. The signal tone was integrated into the dialog as a WAV file and attached to the WoZ tool. The same applied to the dialog task, which was transferred as a picture to the HU. After the participant (blue) had asked a question to the simulated SDS according to the dialog task, the wizard decided whether all requirements were met. If this was not the case, the participant was asked to repeat his utterance. If the participant's question contained all relevant information (*i.e.*, type of question and vehicle function), the participant received an explanatory answer. Just like the signal tone, system-side voice output in this study was synthesized in advance using TTS synthesis (Nuance Vocalizer Studio 3.0.2[8], female voice: Petra-ML) and played back as a WAV file in the WoZ tool. After a pause of three seconds, in order not to expose the participants to additional time pressure, they were asked to rate the previously heard answer. For this purpose, a male voice (Nuance's Yannick-ML) was used at this point so that the participants could easily distinguish the survey from the actual dialog interaction. The participant's rating was then entered in the WoZ tool by the experimenter. After another pause of ten seconds, this procedure was repeated a total of 16 times per subject to assess all combinations of sentence type, domain and driving complexity (s. Figure 3.7). The information about these system parameter types and additional data concerning the dialog flow (grey markers in Figure 3.8) were logged by the WoZ tool.

3.2.1.3.3 Experimental Setup. The study was conducted in a fixed-base driving simulator at the Daimler site in Ulm, Germany. The driving simulator consisted of a halved version of a Mercedes-Benz S-class vehicle (s. Figure 3.9).

Figure 3.10 illustrates the setup of the test environment. While the participant sat in the driver's seat on the driver's side, the responsible experimenter was seated at a table behind the driving simulator. Through the plexiglass window at the back of the driving simulator, it was possible to observe the participant while driving. In this position, the experimenter

[8] `https://www.nuance.com/de-de/omni-channel-customer-engagement/voice-and-ivr/text-to-speech/vocalizer.html` (Online 12/09/2020)

This figure has been removed due to copyright limitations.

Figure 3.9: The driving simulation setup at the site in Ulm with three screens.

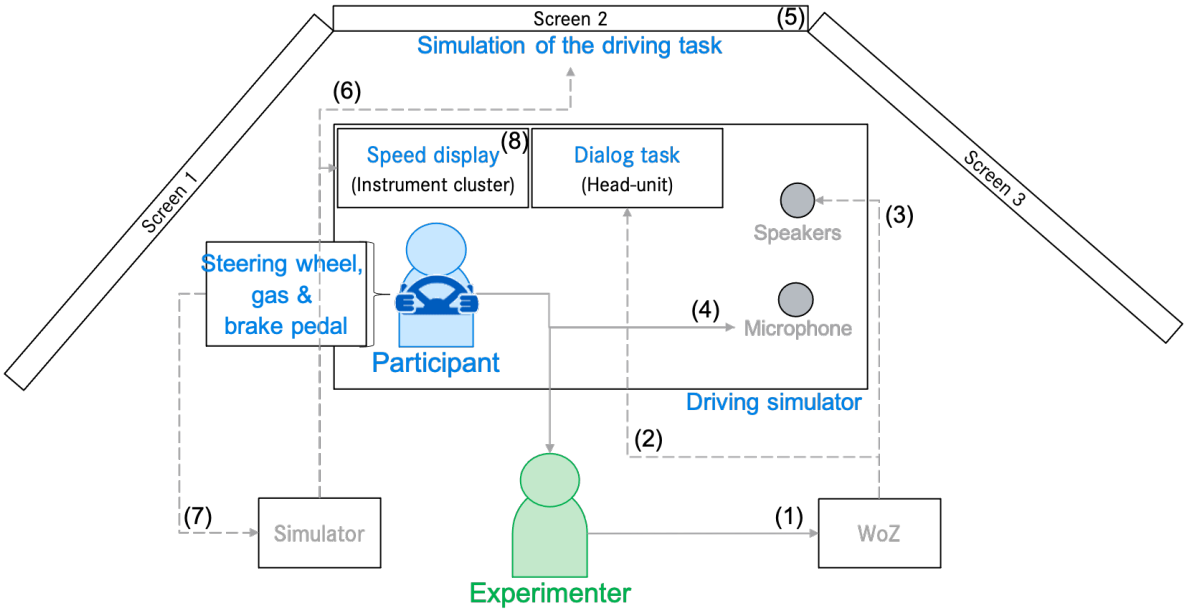


Figure 3.10: Components of the test environment.

controlled the WoZ tool (1) and thus the experiment and dialog flow. On the HU screen in the center console, the dialog task to be performed was displayed to the participant (2). The corresponding voice output of the WoZ was fed to the participant via the loudspeakers installed in the driving simulator (3). The microphone (4) on the sun visor, which was not visible to the participant, enabled the experimenter to follow the participant's verbal utterances. The driving simulator was positioned in front of three screens (5), onto which the driving simulation (6) by means of the Virtual Test Drive toolset by Vires[9] was projected via three projectors. During the study, the participant manually operated the steering wheel, accelerator and brake pedal (7) in the manual driving mode MAN according to the vehicle position on the simulation track. A speedometer was projected on the instrument cluster screen above the steering wheel (8), so that the participant could control his or her own driving speed.

To change the driving complexity, the simulation and track were manually reset by the experimenter. In the case of AUT, a manual driving mode was started, which did not support interventions of the participant in the driving behaviour.

3.2.1.4 Procedure

The study was divided into two phases and lasted 45-60 minutes per subject. In the following, the two phases will be outlined in order to give a complete overview of the study procedure.

3.2.1.4.1 Phase 1: Pre-Survey and Instructions. Prior to the start of the study, each participant was asked to sign a declaration of consent to the collection of personal data and recording of sound material, as well as a non-disclosure agreement. Subsequently, the preliminary questionnaire and the questionnaires to assess Techical Affinity and Big Five personality traits were used to collect demographic and personal data. The content of the study was then explained to the subjects (Appendix A.3.2). They were instructed to interact with a voice assistant during a day's drive on a freeway under different driving conditions (AUT and MAN). For MAN, the participants were asked to maintain a constant speed of 100 km/h, to only use the right lane and not to overtake. In order to ensure equal preparation, the spoken dialog procedure in the form of QAS and one dialog task example were explained to the participants. By means of this example, (s. Figure 3.11), they were shown that each dialog task included a visual representation of a COP or DAS function, its full name and the question type What. In

[9] `https://www.mscsoftware.com/de/virtual-test-drive` (Online: 12/09/2020)

This figure has been removed due to copyright limitations.

Figure 3.11: During the driving simulation, the participant was presented with dialog tasks in the HU screen, indicating a question type and a vehicle function. In this example the participant was expected to formulate the question *Was ist der Brems-Assistent?* (eng. “What is the Brake Assist?”).

addition, the subjects were familiarized with the evaluation scale to be used (Appendix A.3.3). More precisely, they were asked to assess explanatory voice prompts according to their perceived naturalness and comprehensibility jointly on a 5-point Likert scale. In order to ensure a uniform understanding of the concepts naturalness and comprehensibility, they were introduced to concrete evaluation criteria. Additionally, the participants were explicitly instructed to focus on the quality and formulation of prompts and to ignore marginal aspects, such as the TTS quality.

Each participant received a brief introduction to the vehicle controls before the study began.

3.2.1.4.2 Phase 2: WoZ Experiment. A free-way with moderate traffic and a predetermined speed of 100 km/h were chosen as the driving environment. The QAS type What and the manually created syntactic paraphrases (s. Section 3.1) constituted the basis of spoken in-vehicle interactions between participants and the simulated SDS. The procedure of the WoZ experiment is visualized in Figure 3.12: While driving, dialog tasks were displayed in the HU (1; s. Figure 3.11) after an accoustic signal, which indicated the type of question (*i.e.*, What) the subject should formulate to start a QAS and the DAS or COP function they should enquire (s. Figure 3.7). The dialog tasks as presented in the user study can be found in Appendix A.3.4. To keep up the illusion of a real SDS, participants were instructed to activate the voice assistant by saying *Hallo Mercedes* (eng. “Hello Mercedes”) before stating their question (2), for example *Was ist der Brems-Assistent?* (eng. “What is the Brake Assist?”). The simulated SDS subsequently provided a synthesized answer (3). Finally, the participants were asked to evaluate the recently heard answer (4). This procedure was repeated for AUT and MAN. The order of system parameters (driving complexities, domains, sentence types) were randomized across participants.

To further simulate a particular information need and to trigger the participants’ sincere

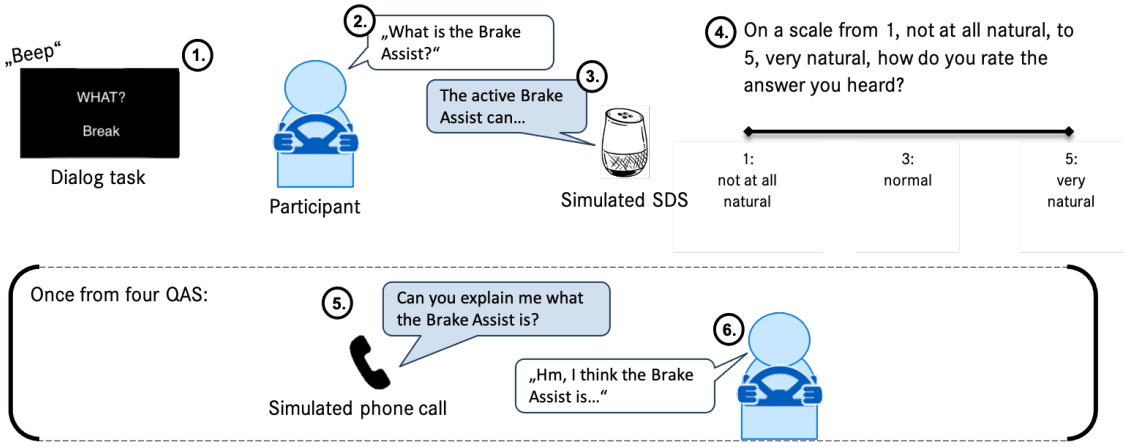


Figure 3.12: Procedure of the WoZ experiment.

interest in the QAS, short test questions in the style of the evaluation survey (*i.e.*, with a male voice for a clear separation from the SDS interaction) were used. They were placed randomly as simulated telephone calls (5) in which the participants were asked to reproduce the previously heard explanation in its general sense (6).

3.2.1.5 Dependent Variables

Against the background of the explorative character of this user study, different types of data were collected to answer the formulated research questions RQ 1 and RQ 2. For this purpose, the WoZ tool produced log files during the simulated SDS interaction from which the entire speech dialog could be reconstructed. As such, for each QAS the respective driving complexity (AUT or MAN), domain (COP or DAS), vehicle function and sentence type (FCV, MCV, NCV or RCV) was stored in addition to the individual evaluation of the participant, which was noted in the WoZ tool by the experimenter. Furthermore, personal data was collected for each participant by means of several questionnaires (s. Subsection 3.2.1.3.1). The collected data was included as variables into the subsequent evaluation.

The experiment was conducted according to the described methodology. The following section will present the results, before they are discussed under consideration of the formulated research questions (s. Subsection 3.2.1.2).

3.2.2 Statistical Analyses and Results

In the following, the most relevant results are presented, starting with the individual questionnaires and followed by the evaluation of the driving simulator data. A summary of results is provided in Appendix A.3.5. Asterisks are employed to indicate if a comparison of conditions was found to be significant. For this purpose, the number of asterisks indicates the level of statistical significance (* $p < .05$, ** $p < .01$, *** $p < .001$).

3.2.2.1 Questionnaire Results

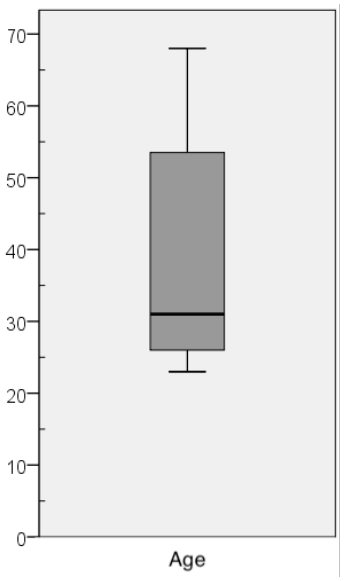
Demographic and personal data was collected by means of several questionnaires (s. Subsection 3.2.1.3.1; each on a 5-point Likert scale).

In total 36 German native speakers participated in the experiment.[10] They comprised 26 male and 10 female participants with an average age of 38.44 years ($SD = 14.63$; s. Figure 3.13a) within a range from 23 to 68 years ($Mdn = 31$ years). The participants considered their linguistic knowledge to be average ($M = 3.19$, $Mdn = 3$, $SD = 1.15$, $IQR = 1.75$; s. Figure 3.13b). Similarly, they indicated an average level of prior knowledge for DAS ($M = 2.92$, $Mdn = 3$, $SD = 1.50$, $IQR = 3$), while they homogeneously reported very little experience with COP ($M = 1.28$, $Mdn = 1$, $SD = 0.61$).

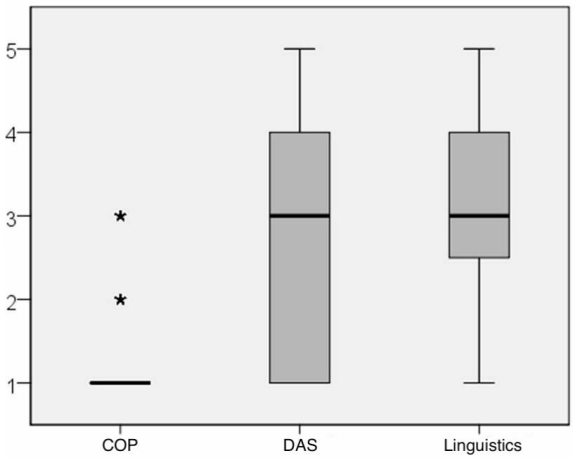
On average, the participants rated themselves as intermediately technically affine ($M = 3.32$, $Mdn = 3.38$, $SD = 0.34$, $IQR = 0.42$). When inspecting the individual Technical Affinity components according to Karrer *et al.* (2009; s. Figure 3.14), the participants rated themselves homogeneously as generally competent (Competence: $M = 3.31$, $Mdn = 3.25$, $SD = 0.48$, $IQR = 0.44$), enthusiastic (Enthusiasm: $M = 3.53$, $Mdn = 3.80$, $SD = 0.87$, $IQR = 1.00$) and rather positive (Positive Attitude: $M = 2.51$, $Mdn = 2.60$, $SD = 0.45$, $IQR = 0.60$) about technical equipment. However, they also indicated a critical view of them (Negative Attitude: $M = 3.91$, $Mdn = 4.00$, $SD = 0.53$, $IQR = 0.95$).

The participants' self-assessed Big Five personality traits according to Rammstedt and Danner (2016) displayed an even distribution (s. Figure 3.15). On average, the participants rated themselves as unselfish and accommodating (Agreeableness: $M = 3.83$, $Mdn = 3.85$, $SD = 0.41$, $IQR = 0.50$) as well as orderly and disciplined (Conscientiousness: $M = 3.88$,

[10] Twelve participants had to be excluded from analyses due to technical problems during the experiment.



(a) Box plot regarding age. The median corresponds to the bold, vertical band in the center of the box.



(b) Box plot regarding prior knowledge in the domains COP, DAS and Linguistics on a 5-point Likert scale.

Figure 3.13: Results of the first user study concerning the age and prior knowledge.

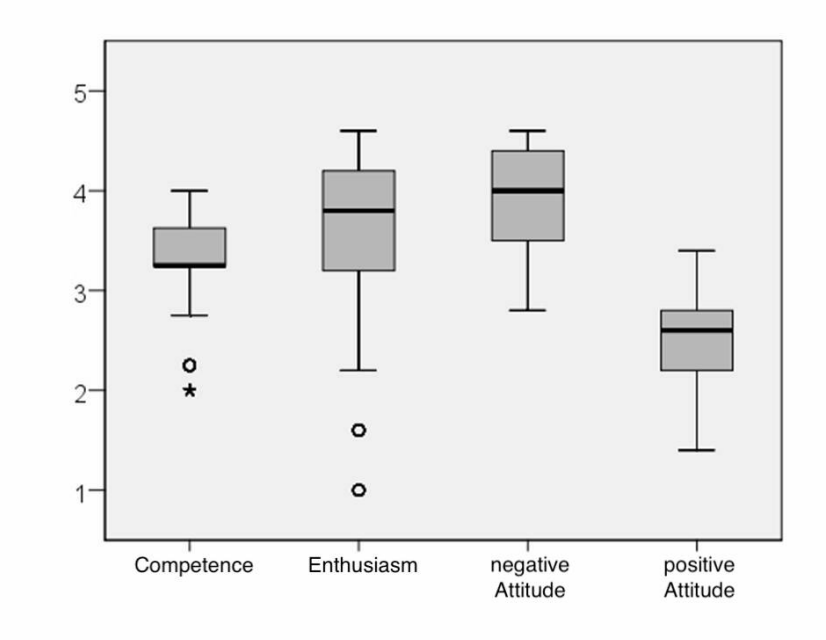


Figure 3.14: Box plot regarding the individual components indicating Technical Affinity according to Karrer *et al.* (2009) on a 5-point Likert scale.

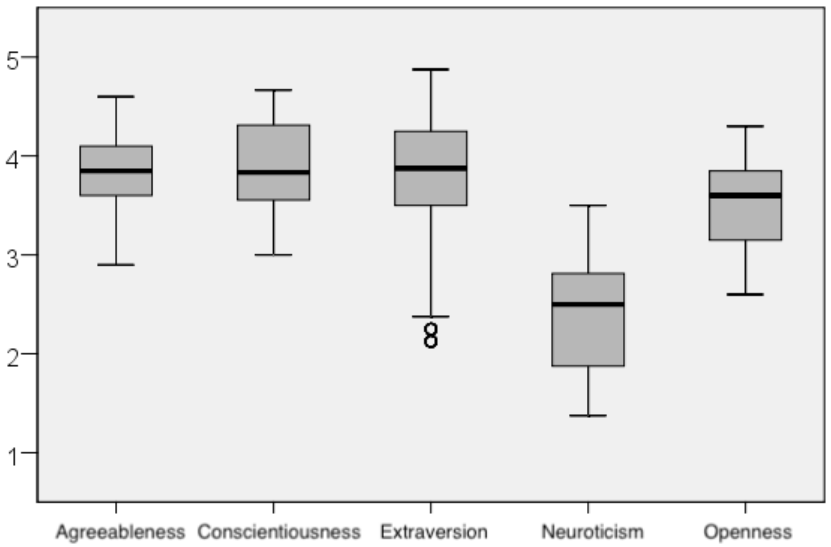


Figure 3.15: Box plot regarding the individual Big Five personality traits according to Rammstedt and Danner (2016) on a 5-point Likert scale.

$Mdn = 3.83$, $SD = 0.50$, $IQR = 0.80$). They indicated that they were generally open to new experiences (Openness: $M = 3.52$, $Mdn = 3.60$, $SD = 0.44$, $IQR = 0.75$) and extraverted (Extraversion: $M = 3.76$, $Mdn = 3.88$, $SD = 0.65$, $IQR = 0.81$). In the case of the latter trait, the largest span of self-reports was observed with a range of 2.75. However, the overall greatest variability in the self-assessments was evident in the case of Neuroticism with an $IQR = 0.97$. In total, the participants rated themselves as low to average neurotic (Neuroticism: $M = 2.37$, $Mdn = 2.50$, $SD = 0.55$).

3.2.2.2 Experimental Results

In the experiment, participants were asked to repeatedly evaluate explanatory voice output realized with different syntactic forms concerning their perceived naturalness and comprehensibility jointly on a 5-point Likert scale.[11] In total, the syntactic prompt paraphrases were assessed 576 times (16 times per participant). As indicated in Figure 3.16, the user ratings were found within the higher ranges of the scale ($M = 3.90$, $Mdn = 4$, $SD = 0.91$, $IQR =$

[11] For reasons of readability, the aspect of comprehensibility will be omitted in the following. Instead, only the aspect of naturalness will be referred to, in which comprehensibility is intended to be implicitly included.

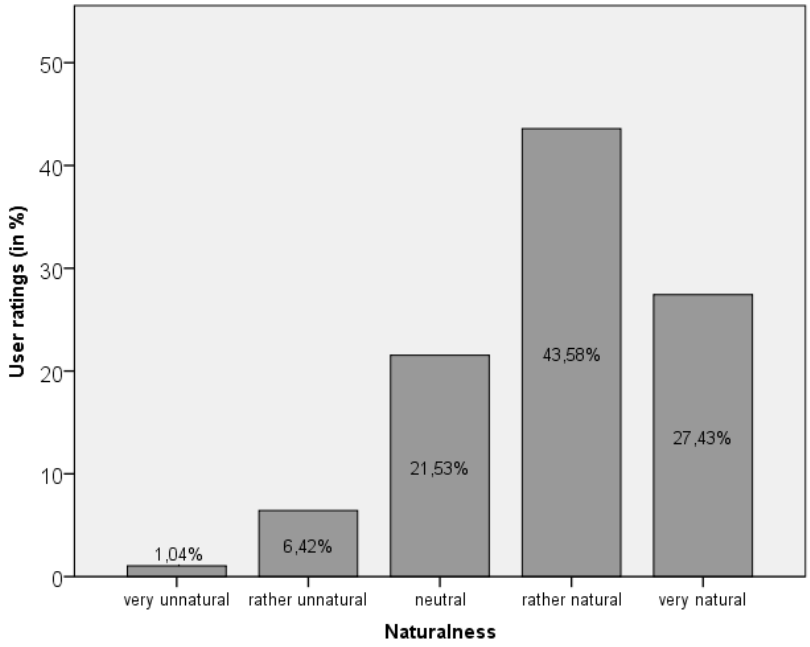


Figure 3.16: Summary of user ratings concerning the perceived naturalness. Adapted from Stier and Sigloch (2019, Figure 2) with kind permission from Association for Computing Machinery.

2), with 43.58% perceived as rather natural. Only 7.46% were assessed as very unnatural or rather unnatural.

In the present study, a repeated-measures design was used. For this reason, further evaluations were conducted fitting a two-level generalized linear mixed model (GLMM, `GENLINMIXED` procedure in `SPSS` v24.0; s. Appendix A.3.5.1) with subjects introduced as random intercepts to account for the repeated-measures character of the data (*cf.* Heck *et al.*, 2013). A cumulative logit link function was chosen given the ordinal scale of the dependent variable $Naturalness$. Thus, in the results below the odds ratio (OR) is used as a measurement to represent the probability of an outcome to occur in the presence of a given condition compared to the probability of the outcome to occur in the absence of that condition. A 95% confidence interval (CI) shown in square brackets is employed to estimate the precision of an odds ratio. The sequential Bonferroni correction was applied for multiple comparisons in post hoc analyses to adjust the computed p-values. Overall, the parameters in Table 3.10 were entered

Table 3.10: Different user and system parameters and their respective levels were entered as fixed effects into a two-level GLMM with the dependent variable *Naturalness* (left). A number of significant main effects were observed (right). Adapted from Stier and Sigloch (2019, Table 3) with kind permission from Association for Computing Machinery.

	Parameter	**Levels**	*Naturalness*
	Complexity	AUT, MAN	*n.s.*
	Domain	COP, DAS	$F(1,523) = 21.686$ ***
	Sentence type (metric)	MCV $<$ FCV $<$ NCV $<$ RCV	$F(1,523) = 8.962$ **
	Age	18-29, 30-44, 45-59, 60-70	$F(3,523) = 5.444$ **
	Gender	female, male	n.s.
Exp[a]	*COP*	low $<$ mid $<$ high	$F(2,523) = 17.436$ ***
Exp[a]	*DAS*	low $<$ mid $<$ high	$F(2,523) = 4.038$ *
Exp[a]	*Linguistics*	low $<$ mid $<$ high	$F(2,523) = 4.012$ *
BFT[b]	*Openness*	mid $<$ high	*n.s.*
BFT[b]	*Conscientiousness*	mid $<$ high	$F(1,523) = 8.853$ **
BFT[b]	*Extraversion*	mid $<$ high	*n.s.*
BFT[b]	*Agreeableness*	mid $<$ high	$F(1,523) = 4.382$ *
BFT[b]	*Neuroticism*	low $<$ mid $<$ high	$F(2,523) = 32.862$ ***
TA[c]	*Competence*	mid $<$ high	*n.s.*
TA[c]	*Neg. attitude*	mid $<$ high	*n.s.*
TA[c]	*Enthusiasm*	low $<$ mid $<$ high	$F(2,523) = 7.632$ **
TA[c]	*Pos. attitude*	low $<$ mid $<$ high	*n.s.*

[a] Experience, [b] Big Five Traits, [c] Technical Affinity
Probability distribution: multinomial; link function: logit (cumulative).
Note: *$p < .05$, **$p < .01$, ***$p < .001$; *n.s.* not significant

as fixed effects into the model. For the purpose of better interpretability, the values of metric and ordinal variables were recoded into subgroups. As such, the participants were divided into four groups (18-29, 30-44, 45-59, 60-70) according to their indicated age. Additionally, each parameter concerning prior experiences, Big Five traits and Technical Affinity components was divided into a maximum of three sublevels, representing a low, mid or high degree in the respective variable. An overview of this procedure is provided in Appendix A.3.5.2.

3.2.2.2.1 Research Question 1: Influencing Factors on the Perception of Voice Output. For the purpose to answer RQ 1, simple main effects were considered in order to identify the influence of particular user and system parameters on the perception of voice output in general. As summarized on the right side of Table 3.10, a number of the parameters proved to be influencing factors.

Age. Among the user-related parameters, the membership to an Age group revealed a statistically significant effect on the prediction of whether a voice prompt was rated as natural (F(3,523)= 5.444, $p < .001$). Overall, the odds of participants aged between 60 to 70 considering voice prompts as natural was 0.188 [0.029, 1.241] times higher than that of 18-29 years old ($p = .083$), 0.043 [0.009, 0.204] times higher than that of 30-44 years old ($p < .001$) and 0.030 [0.002, 0.355] times higher than that of 45-59 years old subjects ($p = .006$). No clear difference in the perception of $Naturalness$ between 18-59 years old participants was observed.

Prior Experience. Concerning prior experiences, all parameters indicated significant differences between their respective levels. As for $Linguistics$ (F(2,523)= 4.012, $p = .019$), the participant group with the lowest self-assessed values indicated a 0.186 [0.052,0.673] and 0.057 [0.005, 0.605] times higher odds ratio of rating voice prompts as natural compared to subjects with average linguistic experience ($p = .010$) or high linguistic experience ($p = .018$). At the same time, no statistically significant difference in the assessment of participants with an indicated ordinary linguistic knowledge was found compared to highly experienced subjects. In the case of DAS (F(2,523)= 4.038, $p = .018$), participants with an indicated average prior knowledge were more likely to rate voice prompts as natural in comparison to low experienced (OR 13.228 [2.079, 84.177]; $p = .006$) or highly experienced (OR 10.201 [1.375, 75.675]; $p = .023$) subjects. A comparison between participants with high and low prior knowledge did not reveal a statistically significant result. Similarly, for COP (F(2,523)= 17.436, $p < .001$) a higher probability to perceive voice output as natural was found for participants with ordinary experience compared to the subgroups with more experience (OR 180.909 [22.666, 1443.918]; $p < .001$) or less experience (OR 24.095 [6.668, 87.068]; $p < .001$). At the same time, participants with a low prior knowledge concerning COP were more likely to assess voice prompts as natural than participants with an indicated high experience (OR 0.133 [0.018, 0.979]; $p = .048$).

Big Five Traits. Overall, user personality was identified as a parameter related to the perception of voice prompts. More precisely, significant differences were found between the different

levels of the Big Five trait *Conscientiousness* (F(1,523)= 8.853, $p = .003$). Here, the odds ratio to rate voice prompts as natural was 0.116 [0.028, 0.480] times higher for highly conscientious participants compared to lower manifestations of this trait. The same was observed for *Agreeableness* (F(1,523)= 4.382, $p = .037$), where highly agreeable participants revealed a higher probability to perceive voice output as natural than participants assigned to the lower level (OR 0.221 [0.054, 0.911]). In the case of *Neuroticism* (F(2,523)= 32.862, $p < .001$), low neurotic (OR 0.067 [0.013, 0.343]; $p < .001$) and highly neurotic (OR 0.001 [0.000, 0.005]; $p < .001$) participants were more likely to rate voice prompts as natural than participants with an average manifestation of this trait. Simultaneously, voice prompts were more likely to be assessed as natural by highly neurotic participants than by subjects assigned to the lowest level (OR 0.011 [0.001, 0.103]; $p < .001$).

Technical Affinity. Finally, also a relation between Technical Affinity and the perceived *Naturalness* of voice output was found. In particular, it was observed that as the level of *Enthusiasm* increased, the odds of evaluating voice output as natural decreased. As such, highly enthusiastic participants were less likely to assess voice prompts as natural compared to participants assigned to the average level (OR 6.930 [1.348, 35.632]; $p = .021$) or lowest level (OR 142.939 [11.740, 1740.320]; $p < .001$). At the same time, the odds of low enthusiastic participants to rate voice output as natural was 0.048 [0.006, 0.400] times higher than for ordinary enthusiasts ($p = .005$).

Domain. Among the system parameters, a significant differences became apparent for the *Domain* (F(1,523)= 21.686, $p < .001$). The odds ratio of voice prompts being rated in a higher category of the dependent variable *Naturalness* for DAS was revealed 0.174 [0.083, 0.363] times that of COP. Thus, voice prompts on DAS were more likely to be assessed as natural compared to prompts on COP.

Sentence Type. One main aspect of this investigation was to identify the role of syntactic forms and *Sentence types* in voice output in a driving context. Here, a significant difference became apparent for the perceived naturalness of syntactic structures (F(1,523)= 8.962, $p = .003$). It was furthermore suggested that an increase in the complexity of syntactic structures was associated with an increase in the odds ratio (OR 1.701 [0.805, 3.594]) of a prompt being perceived as natural. However, this comparison was found to be insignificant ($p = .163$), thus there was no clear difference in the perceived *Naturalness* of the *Sentence types* on an individual level. In addition to the explorative GLMM approach, Wilcoxon signed-rank tests confirmed this observation.

Table 3.11: *Post hoc* analyses for significant interaction effects with $Sentence\ type$ based on the dependent variable $Naturalness$.

Parameter levels		***Naturalness*** Odds ratio [95% CI]			**Interpretation:** MCV ⟵ preference ⟶ RCV				
Age	18-29 vs. 30-44	1.536	[1.222, 1.931]	***	18-29	–	45-59 – 30-44	–	60-70
	18-29 vs. 45-59	1.473	[1.072, 2.022]	*					
	18-29 vs. 60-70	1.751	[1.280, 2.396]	***					
	30-44 vs. 45-59	0.959	[0.742, 1.238]						
	30-44 vs. 60-70	1.140	[0.840, 1.546]						
	45-59 vs. 60-70	0.841	[0.597, 1.184]						
DAS	low vs. mid	0.546	[0.397, 0.750]	***	mid	–	high	–	low
	low vs. high	0.700	[0.494, 0.990]	*					
	high vs. mid	0.780	[0.551, 1.104]						
Extra.	high vs. mid	0.743	[0.574, 0.961]	*	mid	–	–	–	high
Neuro.	low vs. mid	2.034	[1.353, 3.057]	***	low	–	high	–	mid
	low vs. high	1.274	[0.665, 2.441]						
	high vs. mid	1.597	[1.024, 2.489]	*					
Compet.	high vs. mid	0.661	[0.527, 0.830]	***	mid	–	–	–	high
Neg. Att.	high vs. mid	0.718	[0.568, 0.907]	**	mid	–	–	–	high

Probability distribution: multinomial; link function: logit (cumulative).
Comparisons are based on the first named parameter level and MCV as referent.
Note: *$p < .05$, **$p < .01$, ***$p < .001$

3.2.2.2.2 Research Question 2: Influencing Factors on the Perception of Syntactic Forms.

The interactions of all fixed effects listed in Table 3.10 with the parameter $Sentence\ type$ were investigated to answer RQ 2 and to identify which parameters influence the perception of syntactically differing voice prompts. The most important results are summarized in Table 3.11.

Age. In this context, a statistically significant difference between Age groups was demonstrated (F(3,523)= 6.210, $p < .001$). This effect can be explained by a lower probability for the youngest age group between 18 and 29 years to rate voice prompts with an increasing syntactic complexity as natural compared to 30-44 years old participants (OR 1.536 [1.222, 1.931]), 45-59 years old participants (OR 1.473 [1.072, 2.022]), and 60-70 years old participants (OR 1.751 [1.280, 2.396]). No further significant differences between the remaining age groups were observed.

Prior Experience. Similarly, the interaction between $Sentence\ type$ and prior experiences with the domain DAS revealed statistically significant differences (F(2,523)= 7.057, $p < .001$). More precisely, low experienced participants were more likely to assess an increasing syntactic complexity as natural compared to participants with a higher knowledge (mid: OR 0.546 [0.397, 0.750]; high: OR 0.700 [0.494, 0.990]) in this domain. No clear difference was found between the participants with average and high DAS experience.

Big Five Traits. Regarding the interaction effect with the Big Five trait $Extraversion$ (F(1,523)= 5.143, $p = .024$), strong extraverts appeared to rate complex syntactic variants as more natural with a higher probability than less extraverted participants (OR 0.743 [0.574, 0.961]). Meanwhile for $Neuroticism$ (F(2,523)= 9.421, $p < .001$), the participants who assessed themselves as ordinarily neurotic showed odds ratios 2.034 [1.353, 3.057] and 1.597 [1.024, 2.489] times higher than low and highly neurotic subjects, respectively. Thus, they were more likely to assess voice prompts with an increasing syntactic complexity as natural, while no clear difference between low and highly neurotic participants was observed.

Technical Affinity. Furthermore, the inspection of the interaction between Technical Affinity components and the parameter $Sentence\ type$ revealed statistically significant differences. As such, for both components $Competence$ (F(1,523)= 12.739, $p =.000$) and $Negative\ Attitude$ (F(1,523)= 7.756, $p = .006$) the probability of rating complex syntactic forms as natural was higher for high levels of these traits compared to less prominent manifestations ($Competence$: 0.661 [0.527, 0.830]; $Neg.\ Attitude$: 0.718 [0.568, 0.907]).

3.2.3 Discussion of Results and Reflections on the Study Design

In this section, the results of the driving simulator study are discussed against the background of the formulated research questions RQ 1 and RQ 2.

In this user study, several user and system parameters were identified to be related with the perceived naturalness and comprehensibility of in-vehicle voice prompts. As such, it became apparent that familiar domains like DAS, where a functionality can particularly be derived from its name, are more likely to be perceived as natural and comprehensible than unpopular content like COP. Similarly, the age of SDS users appears to be related with their perception of in-vehicle voice output. Here, the oldest group of 60- to 70-year-old participants in particular was identified with a greater probability of rating voice output as natural and comprehensible than younger age groups. In contrast, the younger groups surface as more critical in their evaluations. The aspect of apparently critical users is also found in other parameters, such as in the results concerning the influence of previous experience. The results of this user study indicate that voice output is assessed as less natural and less comprehensible with increasing levels of prior knowledge in COP, DAS and Linguistics. This finding is consistent with the observation that the more enthusiastic and open participants are about technical devices, the more critical and less natural voice prompts are rated. Consequently, the probability of perceiving voice output as natural and comprehensible increases with a decreasing level of prior knowledge and general enthusiasm for technical devices. In this study, it is additionally observed that individual Big Five traits are a relevant component in the perception and evaluation of voice output. In particular, the different manifestations of the personality traits Agreeableness, Conscientiousness, and Neuroticism reveal a clear influence on the evaluation of the perceived naturalness and comprehensibility. Here, with an increasing manifestation of the traits, also the probability of perceiving voice output as natural and comprehensible increases. The observations described here suggest that there are considerable differences in the perception of in-vehicle voice output in dependence of various user and system characteristics. Overall, the results of this user study thus generally confirm the assumption underlying this work that special attention should be paid to the design of voice output, for instance when introducing new domains and functionalities, taking into account individual user characteristics such as age and personality.

The results further indicate that the syntactic form has a clear influence on the perceived naturalness and comprehensibility of voice output. They suggest that as the complexity of syntactic structures increases, so does the perceived naturalness and comprehensibility of voice prompts. However, on an individual level, the participants' ratings in this study for the sentence types did not reveal statistically significant differences. Although there is no evidence to conclude that the individual sentence types differed, they jointly demonstrated an influence on the perceived naturalness and comprehensibility. This has obvious implications for the

design of in-vehicle voice output: The choice of a syntactic form in voice prompts and its inherent complexity is a relevant question and should be related to other system and user parameters in the context of adaptive voice output.

Similar to the observations above, various parameters emerged as relevant influencing factors related to the syntactic complexity of prompts on the perceived naturalness and comprehensibility of voice output. For instance, the youngest age group of 18-29 year olds is less likely to assess voice prompts with an increasing syntactic complexity as natural and comprehensible compared to older age groups. Consequently, older age groups appear to have a different perception of syntactic forms in voice output than younger ones. Similarly, a difference concerning syntactic complexity is observed for different levels of prior knowledge in the domain DAS. While participants with an indicated low experience are more likely to rate complex voice prompts as natural and comprehensible, this probability decreases for participants with higher knowledge in this domain. A similar behavior is apparent for the Technical Affinity components Competence and Negative Attitude. The more competent and critical participants consider themselves with regard to technical devices in genereal, the more likely they are to rate complex voice output as natural and comprehensible. Consequently, voice prompts with a comparably low syntactic complexity are perceived as more comprehensible and natural by participants who indicate a lower manifestation of these characteristics. A similar pattern is revealed for the Big Five traits Extraversion and Neuroticism. The more strongly participants estimate their manifestation in each of these personality traits, the more likely they are to rate complex voice output as natural and comprehensible. Syntactic complexity thus makes a difference in the perception of naturalness and comprehensibility for different manifestations of personality traits. This observation is of particular interest from a linguistic perspective: It is well known that the personality of a person is directly reflected in his or her language behavior (s. Section 2.4). In addition, the results of this study demonstrate that personality also is related with the perception of in-vehicle voice output in different syntactic forms. This finding can be interpreted in the context of a user's own linguistic behavior. Consequently, the question arises whether users prefer a dialog system whose language output reflects their own linguistic behavior and manifests a similar personality (*similarity attraction*, *e.g.*, Nass *et al.*, 1995) – or an opposite personality with deviating linguistic behavior (*complementarity principle*, *e.g.*, Isbister and Nass, 2000; s. Section 2.2).

Overall, the results of this user study suggest the necessity to take real user preferences under consideration of individual system and user characteristics as a basis in the develop-

ment of adaptive in-vehicle SDSs in order to enable a rich user experience. However, at this point, it is necessary to critically question the procedure of the study presented in this section and thus the obtained results: In the course of the user study, a high degree of stress became apparent for the participants during the experiment in the driving simulator. The design of the experiment was intended to create a research environment that was as realistic as possible by having the subjects interact with a simulated dialog system in parallel with the primary task of driving and answering telephone calls in between. However, the time-compressed nature of these demands unintentionally induced an overload of the participants. In order to be able to interpret the obtained results accordingly against this background and to weight their significance, the procedure was subjected to another examination. According to this, the driving task of the participants in the driving simulator was retrospectively classified as highly demanding in the sense of OSPAN (Strayer *et al.*, 2016). In particular, a supply and demand problem of cognitive resources and task performance (Wickens, 2002) was identified caused by the randomly placed, simulated telephone calls and the task of memorizing and reproducing particular contents in combination with further tasks (*i.e.*, driving, keeping a predefined speed, formulating a request, listening to explanatory voice prompts and evaluating them). For instance and taking the results of the Pilot Studies in Sections 3.1.5 and 3.1.6 into account, it seems that the complexity of the driving task might have overshadowed the capability of drivers to subconciously differentiate and prioritize different syntactic forms. Thus, minor syntactic differences were excluded from the perceptile range. For these reasons, the obtained results are considered as a starting point to determine the influence of syntactic forms in voice output and whether user- and system-related parameters should be taken into account in the design of SDS voice output. However, in order to verify the observations presented above, there is a need to conduct a similar experiment clearly focusing on the cognitive demand and concurrent tasks. For this reason, a revision of this user study in consideration of the critical points presented here will be described in Section 3.3.

3.3 Specifying the Influence of Syntax in a Dual-Task Environment

Based on the findings of the user study presented in Section 3.2, a second experiment was conducted. Following the model of the previous study, it took place as a WoZ experiment in

a driving simulator and aimed at specifying the influence of syntactic forms and their inherent complexity on user experience in the context of spoken interaction as a secondary task embedded in one-shot QAS. For this purpose, participants were asked to interact with a simulated SDS and evaluate its voice output in the form of voice prompts with different syntactic complexity. The previous study procedure, however, was revised and the weaknesses outlined in Section 3.2 were resolved.

The following sections are based on the publication by Stier *et al.* (2020b) and provide a detailed description of the employed methodology, before results are presented and discussed.

3.3.1 Methodology

In this section, the methodological approach of the user study is presented. The participants and the chosen experimental design are described, before the employed materials and the procedure get introduced.

3.3.1.1 Participants

In this experiment, a total of 50 German native speakers with an average age of 42.60 years (SD 14.97) and a gender distribution of 30 male and 20 female subjects participated. All of them possessed a valid driver's license and received an expense allowance of 50€ for their participation.

3.3.1.2 Experimental Design

The study design of this experiment was based on the model of the previous study, but also includes some adaptations. An overview is provided in Figure 3.17. As shown here, the perception of voice output was studied in a two (two driving complexity levels AUT and MAN) x two (two domains COP[12] and DAS) x three (three question types What, How, When) x two (two sentence types MCV and RCV) within design. Overall, each participant thereby experienced 24 QAS. The study focused on the perceived naturalness and comprehensibility

[12]The COP function *Freude* ("Joy") has been replaced by the term *Vergnügen* ("Joy"). Furthermore, the DAS function *Nothalt-Assistent* ("Emergency Stop Assist") was replaced by the function *Brems-Assistent* ("Brake Assist"). Instead of the *Brems-Assistent*, the *Nothalt-Assistent* was now used as an example in explanations.

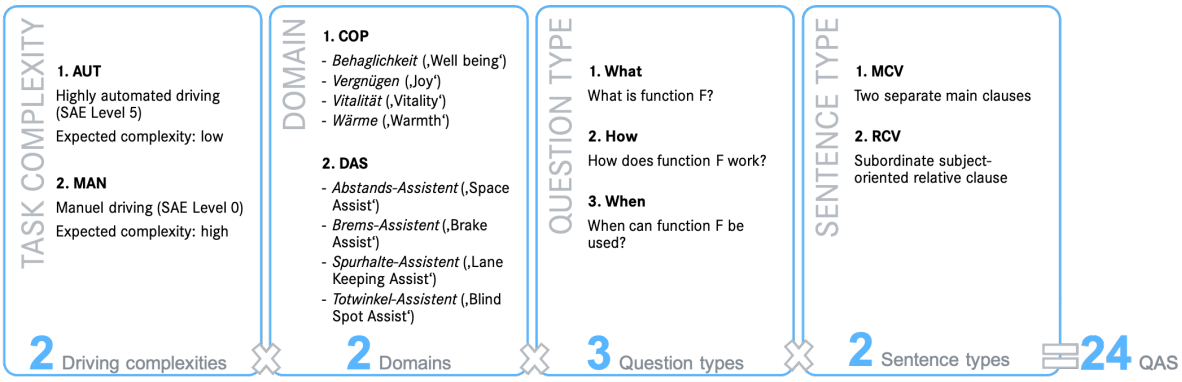


Figure 3.17: The set of parameters considered in the experimental design.

of syntactically different voice prompts. Furthermore, the influence of listening to voice output of different syntactic complexity on driving performance was investigated.

3.3.1.3 Materials

In this subsection, the materials of this user study are described.

3.3.1.3.1 Questionnaires. Following the previous study, different questionnaires were used to collect demographic and personal data from the participants. The applied questionnaires were created using the tool `soSci`[13] and can be found in Appendix A.4.1 and A.4.2.

- **Preliminary Questionnaire:** In this preliminary questionnaire, demographic information (age, gender, etc.) about the participants were collected. In addition, they were asked to self-assess their level of linguistic knowledge and prior experiences with the domains COP and DAS on a 5-point Likert scale.
- **Technical Affinity:** The questionnaire by Karrer *et al.* (2009) covers the four components Competence, Enthusiasm, Negative and Positive Attitude and consists of 19 items, for which a 5-point scale was used. It is a widely used measure to capture the affinity for technology (s. Section 2.4).

[13] `https://www.soscisurvey.de` (Online 01/21/2021)

- **Big Five Personality Traits:** Participants were asked to self-assess their personality traits with the help of the German version of the BFI questionnaire by Rammstedt and Danner (2016) on a 5-point scale. It consists of 45 questions, which are assigned to the five personality traits Agreeableness, Conscientiousness, Extraversion, Neuroticism and Openness. It is a frequently applied instrument to reduce human behavior to a small number of interpretable dimensions (s. Section 2.4).
- **DALI:** In the half and at the end of the user study, the participants were asked to complete the DALI questionnaire based on Hofmann (2015). It is an established method for evaluating the cognitive load of users on the basis of six dimensions (s. Section 2.3). For each dimension, one question was asked on a 5-point Likert scale.

3.3.1.3.2 Wizard-of-Oz as Simulated Spoken Dialog System. In order to simulate spoken interaction between participants and a real SDS, a Daimler internal tool on the model of SUEDE (Klemmer *et al.*, 2000) was used. For the purpose of this experiment, the dialog flow specification of the user study presented in Subection 3.2.1.3.2 was reused and adapted according to the changes in the procedure of the present study. As such, additional dialog tasks in the form of pictures and corresponding WAV files representing explanatory system answers were attached to the WoZ tool according to the Question and Sentence types explained in the following subsection. Similarly, these parameter types were included into the logging information of the WoZ tool for subsequent analyses.

3.3.1.3.3 Question and Sentence Types. While the driving complexity levels and domains of the previous study presented in Section 3.2 were adopted, in this experiment a total of three Question types (What, How and When) and two Sentence types (MCV and RCV) were included as controlled system parameters. The question types are defined as follows:

What As before, the conceptual explanation to the question *Was ist Funktion F?* (eng. “What is function F?”) supplies a general definition of a particular function F. The transfer of factual knowledge in association with a concrete function is focused here.

How The question *Wie funktioniert Funktion F?* (eng. “How does function F work?”) requires an explanation of F’s functionality. The focus here is on the transfer of methodical knowledge in association with a concrete function. It is described how the functional scope

This figure has been removed due to copyright limitations.

Figure 3.18: Exterior view of the driving simulation setup with a 180-degree screen.

described by the question type What is technically realized. An explanation for the question type How thus presupposes the factual knowledge from the conceptual explanation of the question type What.

When The additional question *Wann ist Funktion F einsetzbar?* (eng. "When can I use function F?") represents a special case of How in this work, asking for particular limitations of function F.

The expansion of the range of question types required the re-creation of syntactic paraphrases, relying on the methodological approach described in Section 3.1. The paraphrases that were manually created in this way are listed in Appendix A.4.3.

3.3.1.3.4 Experimental Setup. The study took place in a fixed-base driving simulator of a Mercedes-Benz C-Class (s. Figure 3.18) at the Daimler site in Sindelfingen, Germany.

The test environment is illustrated in Figure 3.19. While the participant sat in the driver's seat on the driver's side, the responsible experimenter sat at a control station and monitored the participant via cameras (behind the HU and on the right A-pillar; 1) and was connected to him/her via an intercom system. This precaution served in particular to prevent motion sickness. At any time, the experimenter was able to communicate with the participant and, if necessary, to interrupt or stop the experiment. In addition, the experimenter observed the ongoing speech dialogs via a microphone installed in the rearview mirror (2) and controlled the WoZ tool (3). The dialog tasks to be performed were displayed on the HU screen (4; s. Figure 3.20), while dialog turns directed by the WoZ were played via a speaker installed in the front passenger footwell (5), not visible to the driver. The driving simulator was positioned in front of a 180-degree screen (6), onto which the driving simulation (7) was projected via four projectors. During the manual driving mode MAN, the participant manually operated the steering wheel, accelerator and brake pedal (8) according to the vehicle position on the simulation track. For this purpose of speed control, a speedometer was projected on the instrument cluster screen (9). Furthermore, Real-Time Driving Data (RTDD) was assessed at intervals of 2 ms via a Controler Area Network (CAN bus). The vehicle bus was additionally

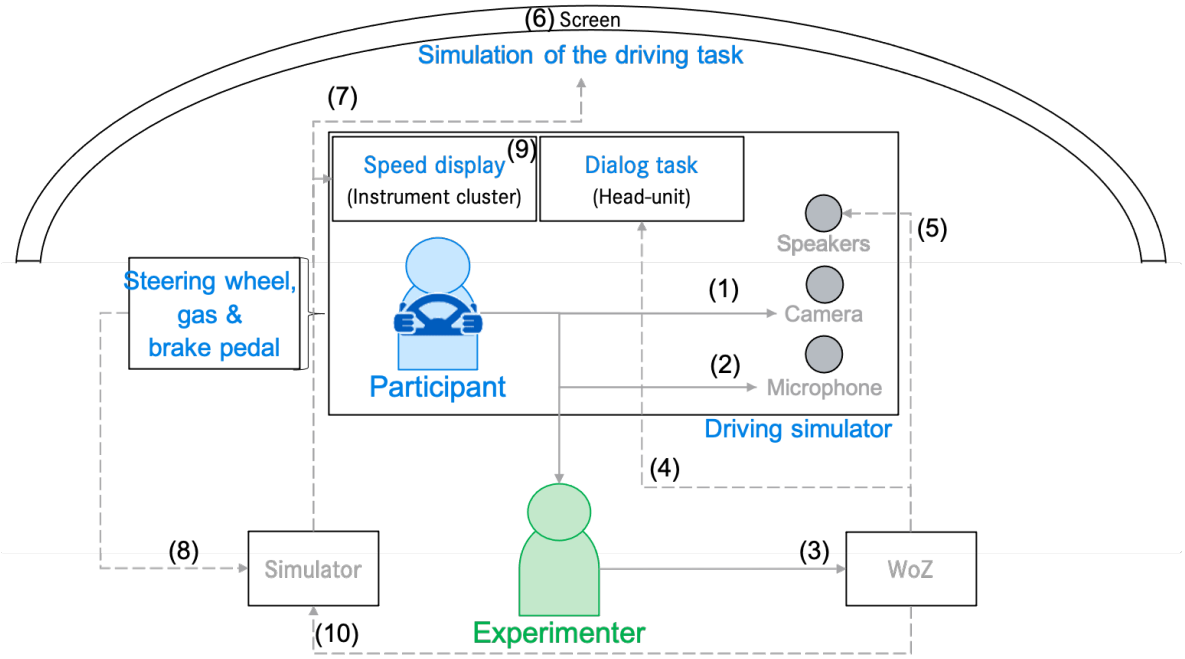


Figure 3.19: Components of the test environment.

connected to the WoZ tool to synchronize assessed user ratings with user driving performance (10).

The simulation and track were manually reset by the experimenter in order to change the driving complexity. For AUT, a manual driving mode was startet, which did not support interventions of the participant.

3.3.1.4 Procedure

The procedure of this study was divided into three parts and lasted approximately 60-90 minutes per participant. The following subsections provide a detailed overview of the individual phases.

3.3.1.4.1 Phase 1: Pre-Survey and Instructions. At the beginning of the study, each participant was asked to sign a consent form for the collection of personal data and data collection

This figure has been removed due to copyright limitations.

Figure 3.20: Interior view with driving task in the HU.

during the study. Subsequently, the pre-survey questionnaire, including the self-assessment questionnaires for Technical Affinity and Big Five personality were presented for completion (s. Appendix A.4.1). Then, the content of the study was explained to the participants (s. Appendix A.4.4). They were instructed to interact with a voice assistant during a daytime drive on a free-way under two different driving conditions (AUT and MAN). During the manual driving condition, the participants were asked to maintain a constant speed of 100 km/h, to use only the right lane and not to overtake. Furthermore, the participants were prepared for the spoken interaction with the voice assistant in form of one-shot QAS in the context of three question types (What, How, When) and two domains (COP and DAS) and subsequent evaluation. For this purpose, the participants received an introduction to the concepts of naturalness and comprehensibility. With regard to comprehensibility, the participants were instructed to assess whether, in their opinion, a voice prompt was immediately and intuitively comprehensible without further thought. Concerning naturalness, the participants were asked to evaluate whether they considered the quality and design of the last heard responses to be pleasant and naturally formulated. They were explicitly instructed not to consider aspects such as the TTS voice and error-free pronunciation in their ratings. In addition, the subjects were familiarized with the evaluation scale on a 5-point Likert scale (s. Appendix A.4.6). By means of dialog task examples, it was explained to them that after each QAS they were asked to rate the comprehensibility and after completion of three consecutive QAS (What, How, When) to rate the naturalness of the voice prompts they heard (s. Appendix A.4.5).

Before the study began, each participant received an introduction to the vehicle controls.

3.3.1.4.2 Phase 2: WoZ Experiment. The drive in the simulator took place on a free-way with moderate traffic and a specified speed of 100 km/h. The route included straight stretches as well as some slight curves, as shown by the course in Figure 3.21. The particpants were instructed to follow a lead vehicle at a distance of approximately 100 m (*i.e.*, two delineator posts). The procedure of the WoZ experiment is visualized in Figure 3.22: Baselines were included at the beginning and end of the drive to gather performance data without SDS in-

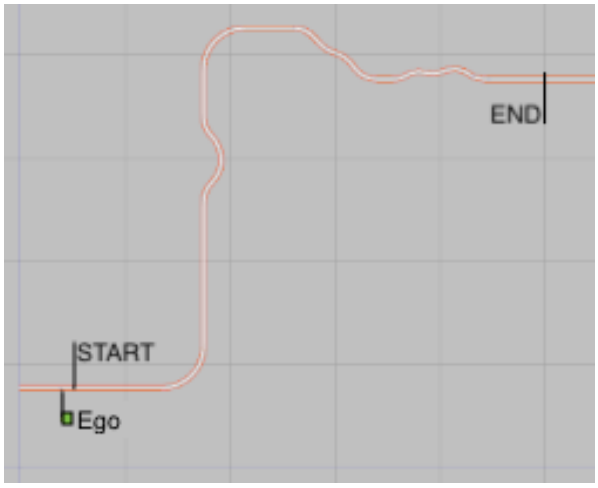


Figure 3.21: Driving simulation route.

teraction as a secondary task. Spoken interaction between participants and the WoZ was based on consecutive QAS for What, How and When. Within a set of these three QAS, the identical sentence type (MCV or RCV) was applied. In each step, user-initiative was triggered by an accoustic signal and displaying a task on the HU screen, indicating the respective type of question to be formulated and a COP or DAS function the participant should enquire. The dialog tasks as presented in the study are provided in Appendix A.4.7. In order to keep up the illusion of real SDS interaction, each participant was instructed to activate the simulated voice assistant with the phrase *Hallo Mercedes* (eng. “Hello Mercedes”) before stating their question. A question by the user was followed by an explanatory voice prompt by the simulated SDS and the request to assess the heard voice prompt concerning its perceived comprehensibility. After completing all three QAS, the participant was asked to evaluate the naturalness of voice output. The outlined procedure was repeated for AUT and MAN. The order of parameters (driving complexities, domains, sentence types) were randomized. Only the sequence of the question types What, How and When was retained.

3.3.1.4.3 Phase 3: Intermediate and Post-Survey. After completing one driving complexity level, the participants were asked to complete an interim DALI questionnaire outside the vehicle (s. Appendix A.4.2). In the meantime, the track and the simulation were restarted by the experimenter. After completing the second drive with the second driving complexity

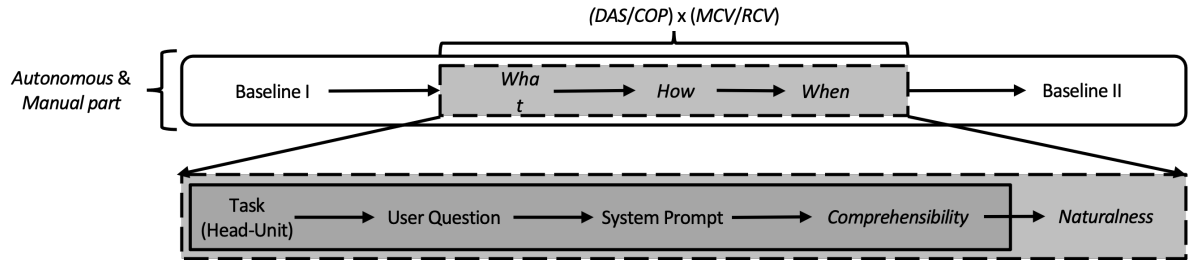


Figure 3.22: Detailed procedure of the WoZ experiment. Taken from Stier *et al.* (2020b, Figure 4).

level, the same DALI questionnaire was to be completed in order to subsequently assess the cognitive load level of both experienced driving complexity conditions.

3.3.1.5 A Modified Study Design

The study design and procedure described above were chosen in order to revise the weaknesses of the user study outlined in Section 3.2. There, the experimental design was retrospectively classified as highly demanding and overloading due to an identified supply and demand problem of cognitive resources and task performance (Wickens, 2002). The difference between the experimental approaches of the present user study and the previous experiment thus mainly consists in the reduction of cognitive load of the originally highly demanding task. For this purpose, certain modifications were made to the study design. As such, a lead vehicle was included as an orientation point of orientation. Thereby, the participants were relieved of the stress factor to maintain speed on their own responsibility. In addition, the short QAS were replaced, which merely consisted of What, with consecutive QAS (What, How, When) in order to prime subjects over a longer time period with a particular syntactic structure. The number of sentence types was reduced from four to two (MCV, RCV) with the most outstanding characteristics and differences concerning their complexity. Thereby, the perceivable differences and the expected effect in a direct comparison of sentence types are expected to increase. Furthermore, the evaluation whether voice output was perceived as comprehensible and natural was split into two separate steps to provide a clearer understanding of the assessment task. Most importantly, the randomly placed phone calls by the WoZ (subjects should reproduce the last heard voice prompt in own words) were omitted to decrease the stress and cognitive

load of participants. The phone calls intended to simulate a sincere interest of the participants in the explanatory voice prompts. Although this motivation is still acknowledged, it has been shown that this approach increased the stress level of participants to an extent that the results should be interpreted with reservations. Finally, both the measurement of RTDD and the inclusion of DALI questionnaires as part of an interim and post-survey were chosen in order to objectively assess the influence of syntactically differing voice output and to ensure different levels of cognitive load induced by two different driving conditions (AUT, MAN) on a subjective level. The modifications described in the experimental design of the present user study are intended to eliminate prior reservations.

3.3.1.6 Dependent Variables

Evaluation measures. In the course of this user study, different types of data were collected to specify the influence of syntactic forms in voice output with respect to user experience and driver distraction. These included the personal data collected in the pre- and intermediate/post-survey and information logged in the driving simulator during the participants' speech interaction with the simulated SDS. As described in Subsection 3.2.1.5, the WoZ tool produced log files from which the speech dialog could be reconstructed. In addition, various driving performance parameters (RTDD) were recorded in parallel for the manual driving part MAN in this user study. Here, the driving speed, distance to the vehicle in front, and lane keeping were measured. No data reflecting user performance was generated during AUT.

Table 3.12 provides an overview of the measures, which were employed based on the collected data. In order to evaluate the influence of syntactic forms in voice output, the perceived naturalness (*Nat*) and comprehensibility (*Comp*) per voice prompt were extracted from the WoZ logs. In order to ensure valid conclusions, the logged driving performance measures were limited to those sequences during which no voice interaction took place (baseline BL) or a voice prompt was played (12/participant in MAN; voice output sequence VOS). For the time intervals of these sequences, standard deviations were computed for the driving speed *SPDev* [in km/h], distance to the lead vehicle *DLDev* [in m] and the lateral position on the lane *LPDev* [in m]. Finally, the DALI questionnaire as part of the intermediate and post-survey provided a subjective assessment of the workload during each driving part.

Hypotheses. On the basis of these evaluation measures, several hypotheses were formu-

Table 3.12: Measures of the user study concerning the evaluation of the influence of syntactic forms in voice output.

	Measure	Source
User experience	Perceived naturalness (*Nat*)	WoZ logs
	Perceived comprehensibility (*Comp*)	WoZ logs
Driver distraction	Speed deviation (*SPDev*)	RTDD logs
	Deviation of distance to lead vehicle (*DLDev*)	RTDD logs
	Deviation of lateral position (*LPDev*)	RTDD logs
	Assessment of workload (*DALI*)	DALI questionnaire (intermediate/ post-survey)

lated. They are presented in the following and will be validated with the statistical analyses results in the following section.

The user study was performed in different driving conditions. In this context, the following fundamental results were expected as a basis for subsequent analyses:

- Driving as a primary task induces a certain cognitive load on the driver. With an increasing degree of automation, this cognitive load is assumed to decrease. Accordingly, it was expected that the manual driving part MAN in the simulator was assessed as more cognitively stressful than the autonomous driving part AUT.

$$DALI_{\text{AUT}} < DALI_{\text{MAN}} \tag{3.3}$$

- During driving, parallel speech-based interaction generally increases a driver's cognitive load. Accordingly, driving performance with respect to the distraction parameters speed, distance to the driver in front, and lane keeping was expected to deteriorate during voice output sequences compared to driving without voice interaction.

$$SPDev_{\text{BL}} < SPDev_{\text{VOS}} \tag{3.4}$$

$$DLDev_{\text{BL}} < DLDev_{\text{VOS}} \tag{3.5}$$

$$LPDev_{\text{BL}} < LPDev_{\text{VOS}} \tag{3.6}$$

When comparing voice output of a differing syntactic complexity, the following results were expected:

- Driving performance is assumed to depend on the syntactic form and its inherent complexity of voice prompts. Since complex syntactic structures are cognitively more demanding than simple structures, an increase in the cognitive load of participants and thereby a deterioration in their driving performance is expected during voice prompts in the form of RCV. A degradation of the driving behavior here includes an increased deviation of the driver distraction parameters speed, distance to the driver in front and lane keeping.

$$SPDev_{\mathrm{MCV}} < SPDev_{\mathrm{RCV}} \tag{3.7}$$

$$DLDev_{\mathrm{MCV}} < DLDev_{\mathrm{RCV}} \tag{3.8}$$

$$LPDev_{\mathrm{MCV}} < LPDev_{\mathrm{RCV}} \tag{3.9}$$

- In general, the perceived naturalness and comprehensibility of voice prompts is expected to be influenced by individual user characteristics and system parameters. In particular, an influence of syntactic forms is expected. Here, voice output in the form of MCV is assumed to be perceived as more comprehensible while driving given a low syntactic complexity. Similarly, MCV is expected to be assessed as more natural and mirroring lifelike linguistic behavior (s. Section 2.2). Analogous to Section 3.2, the relationship between individual system and user parameters, the perception of voice output in general and the influence of different syntactic forms will be specified by means of an exploratory evaluation approach.

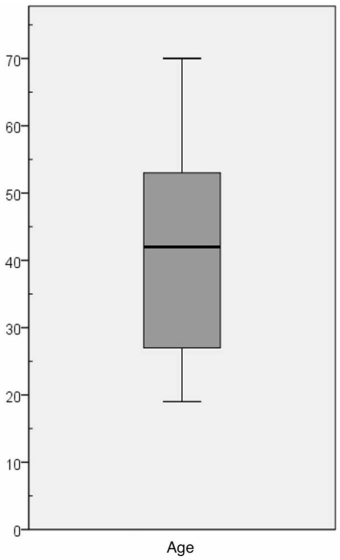
$$Nat_{\mathrm{RCV}} < Nat_{\mathrm{MCV}} \tag{3.10}$$

$$Comp_{\mathrm{RCV}} < Comp_{\mathrm{MCV}} \tag{3.11}$$

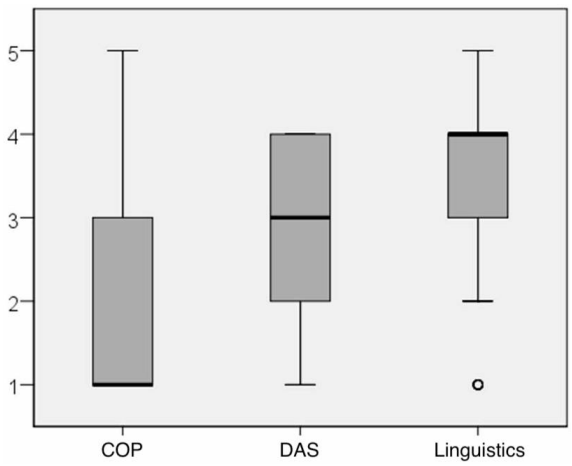
The driving simulation study was conducted according to the described methodology. The following section presents the results of the experiment, before they are discussed and validated against the formulated hypotheses above.

3.3.2 Statistical Analyses and Results

The following subsections present the most relevant results of this user study. First, the results of the applied pre-survey will be described, followed by the evaluation of the subjective assessment of cognitive workload. Second, the results concerning the objective driving performance measures and the investigation of user experience will be presented. A complete



(a) Box plot regarding age. The median corresponds to the bold, vertical band in the center of the box.



(b) Box plot regarding prior knowledge in the domains COP, DAS and Linguistics on a 5-point Likert scale.

Figure 3.23: Results of the first user study concerning the age and prior knowledge.

overview of results is provided in Appendix A.4.8. Asterisks are employed to indicate if and at which level a comparison of conditions was found to be statistically significant (* $p < .05$, ** $p < .01$, *** $p < .001$).

3.3.2.1 Questionnaire Results

As outlined in Subsection 3.3.1.3.1, demographic and personal information of the participants was collected by means of several questionnaires.

Overall, 46 German native speakers participated in the experiment with a gender distribution of 27 male and 19 female subjects.[14] The average age was 41.98 years ($SD = 15.07$) within a range from 19 to 70 years ($Mdn = 42$ years; s. Figure 3.23a).

Further personal information was assessed via questionnaires on 5-point Likert scales. As such, the participants considered their linguistic background as good ($M = 3.89$, $Mdn = 4$,

[14] Four participants were excluded from analyses due to technical problems in the driving simulator.

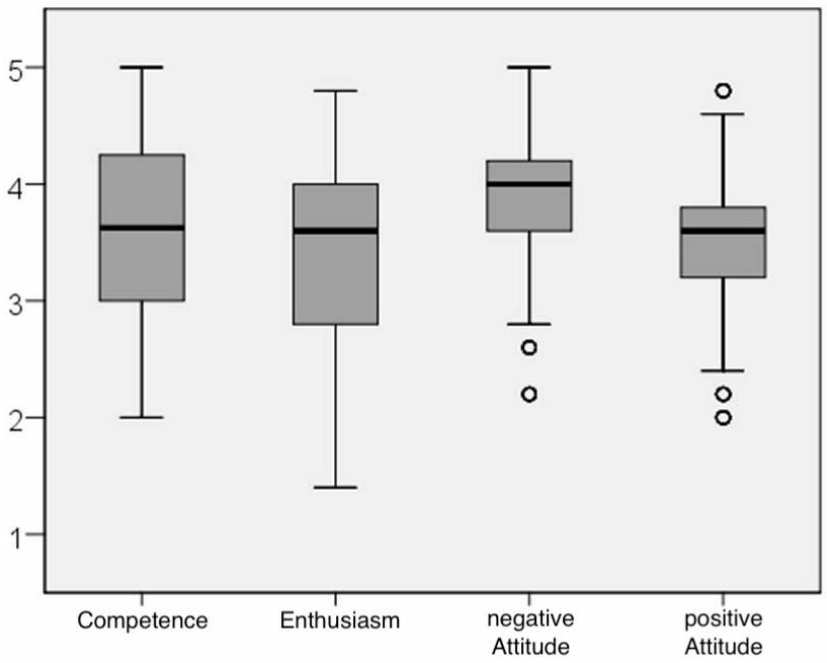


Figure 3.24: Box plot regarding the individual components indicating Technical Affinity according to Karrer *et al.* (2009) on a 5-point Likert scale.

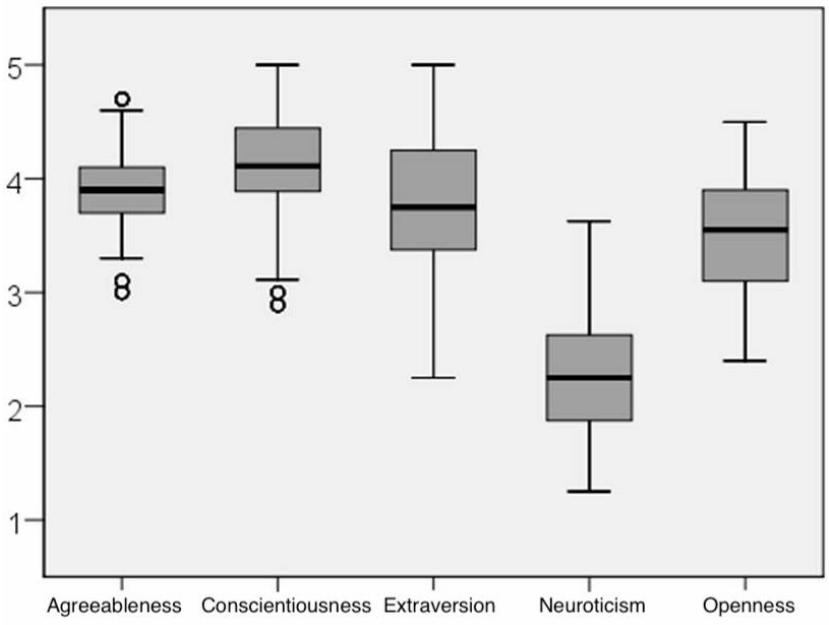


Figure 3.25: Box plot regarding the individual Big Five personality traits according to Rammstedt and Danner (2016) on a 5-point Likert scale.

$SD = 0.87$; s. Figure 3.23b). Similarly, they indicated an average level of prior knowledge in the domain DAS ($M = 2.91$, $Mdn = 3$, $SD = 1.10$). In contrast, the self-assessed experience with the domain COP was very low ($M = 1.83$, $Mdn = 1$, $SD = 1.03$), with reported scores using the entire range of the 5-level scale. Overall, the domain-specific prior experiences appear less homogeneous with $IQR = 2$ each for DAS and COP compared to $IQR = 1$ for Linguistics.

On average, the participants considered themselves as rather technically affine ($M = 3.61$, $Mdn = 3.69$, $SD = 0.53$, $IQR = 0.75$). The individual Technical Affinity components (Karrer *et al.*, 2009) are visualized in Figure 3.24. Here, the participants rated themselves homogeneously as generally competent (Competence: $M = 3.60$, $Mdn = 3.63$, $SD = 0.68$, $IQR = 0.25$), enthusiastic (Enthusiasm: $M = 3.40$, $Mdn = 3.60$, $SD = 0.91$, $IQR = 1.20$) and positive (Positive Attitude: $M = 3.53$, $Mdn = 3.60$, $SD = 0.61$, $IQR = 0.60$) about technical devices – but nonetheless take a critical view of them (Negative Attitude: $M = 3.92$, $Mdn = 4.00$, $SD = 0.57$, $IQR = 0.60$).

Homogeneous self-assessments were observable concerning the participants' self-assessed Big Five personality traits according to Rammstedt and Danner (2016; s. Figure 3.25). Here, the participants assessed themselves as unselfish and tolerable (Agreeableness: $M = 3.92$, $Mdn = 3.90$, $SD = 0.37$, $IQR = 0.40$) as well as orderly and disciplined (Conscientiousness: $M = 4.10$, $Mdn = 4.11$, $SD = 0.49$, $IQR = 0.67$). They furthermore indicated that they were generally open to new experiences (Openness: $M = 3.55$, $Mdn = 3.55$, $SD = 0.52$, $IQR = 0.80$) and extraverted (Extraversion: $M = 3.72$, $Mdn = 3.75$, $SD = 0.68$, $IQR = 0.88$). In addition, the participants considered themselves as low neurotic (Neuroticism: $M = 2.29$, $Mdn = 2.25$, $SD = 0.56$, $IQR = 0.75$).

3.3.2.2 Subjective Assessment of Cognitive Workload

The assessment of the cognitive load was achieved by means of the DALI questionnaire (s. Subsection 3.3.1.3.1). The questionnaire was presented to the participants at two different measurement times, each time after completion of an AUT or MAN drive. In total, the DALI items were recorded twice for each participant on a 5-point Likert scale.

Overall, the participants assessed the cognitive load during AUT ($M = 1.87$, $Mdn = 1.75$, $SD = 0.72$, $IQR = 0.83$) as significantly lower than for MAN ($M = 2.89$, $Mdn = 2.83$, $SD =$

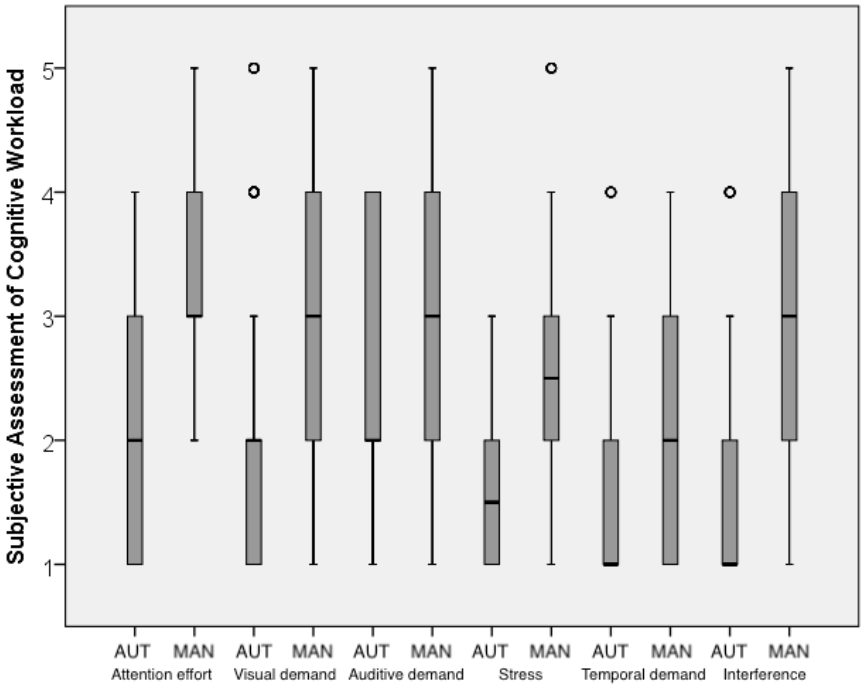


Figure 3.26: Box plots of the individual DALI dimensions on a 5-point Likert scale.

0.69, $IQR = 1.17$; $Z =$ -16.034, $p < .001$, $r = .84$). The results of the individual DALI dimensions are represented in Figure 3.26.

The participants estimated the attentional effort during AUT with a mean value of 2.02 ($Mdn = 2$, $SD = 1.08$) as significantly lower than for MAN ($M = 3.41$, $Mdn = 3$, $SD = 0.80$; $Z =$ -14.972, $p < .001$, $r = .78$). Overall, with an interquartile range of 1, the participants' assessments for MAN were more consistent than in the case of AUT ($IQR = 2$).

Concerning the indicated visual demand, ratings were found more homogeneous in the case of AUT except for two outliners ($M = 1.96$, $Mdn = 2$, $SD = 1.00$, $IRQ = 1$) compared to MAN ($M = 3.24$, $Mdn = 3$, $SD = 0.98$, $IQR = 2$), where the assessments span over the entire scale range. Overall, a significant difference was also observed between AUT and MAN ($Z =$ -13,699, $p < .001$, $r = .71$).

A clear difference became also apparent concerning the auditive demands of AUT ($M =$ 2.48, $Mdn = 2$, $SD = 1.12$) and MAN ($M = 3.02$, $Mdn = 3$, $SD = 1.11$; $Z =$ -10.352,

$p < .001$, $r = .54$). Overall, although the participants' ratings showed a similar variance with $IQR = 2$, they differed in terms of their scale ranges of 3 for AUT and 4 for MAN.

The stress level induced by MAN ($M = 2.63$, $Mdn = 2.50$, $SD = 0.99$, $IQR = 1$) was estimated significantly higher compared to AUT ($M = 1.65$, $Mdn = 1.50$, $SD = 0.73$, $IQR = 1$; $Z =$ -13.268, $p < .001$, $r = .69$).

A similar behavior was observed concerning the indicated temporal demand. In general, the temporal demands in both driving conditions were perceived to be rather low. Here, the ratings for AUT ($M = 1.50$, $Mdn = 1$, $SD = 0.65$, $IQR = 1$) were found to be significantly lower than for MAN ($M = 2.04$, $Mdn = 2$, $SD = 0.98$; $Z =$ -10.608, $p < .001$, $r = .55$), whereby the latter indicated a higher dispersion of ratings with $IQR = 2$.

One of the most striking differences was observed for the dimension interference. Analogous to the previous observations, a significant difference became evident between the ratings for MAN ($M = 3.00$, $Mdn = 3$, $SD = 0.98$) and AUT ($M = 1.59$, $Mdn = 1$, $SD = 0.85$; $Z =$ -14.617, $p < .001$, $r = .76$). While the participants' assessments were rather consistent in the case of AUT ($IQR = 1$), they were found to be spread on the entire 5-point scale ($IQR = 2$).

Overall, the above observations clearly confirm Hypothesis 3.3.

3.3.2.3 Objective Driving Performance Measures

As described in Section 3.3.1.3.4, driving performance measures were assessed during the user study for the manual driving part MAN. No RTDD reflecting user performance was generated in the driving simulator during AUT, thus the analyses of objective performance measures described in this section only refer to MAN. Table 3.13 provides an overview of the results.

In a first step, the driving performance of participants during the combined baseline drives (BL; Baseline I & II in Figure 3.22) was compared with the performance during those driving sequences in which the participants listened to a voice prompt (VOS; including MCV and RCV). All three measures indicated higher values during VOS (*SPDev* 1.029, *DLDev* 13.241, *LPDev* 0.236) compared to BL (*SPDev* 0.809, *DLDev* 5.628, *LPDev* 0.167). Wilcoxon signed-rank tests revealed significant differences in the driving behavior between these two conditions in terms of *SPDev* ($Z =$ -8.344, $p < .001$, $r = .36$), *DLDev* ($Z =$ -18.661, $p < .001$, $r = .79$) and *LPDev* ($Z =$ -15.693, $p < .001$, $r = .67$). Thus, a degradation of the driving behavior from BL to VOS was observed. These findings support the Hypotheses 3.4, 3.5 and 3.6.

Table 3.13: Assessed driving performance measures and the results of Wilcoxon signed-rank tests. Adapted from Stier *et al.* (2020b, Table 4), © 2021 Copyright held by the owner/author(s).

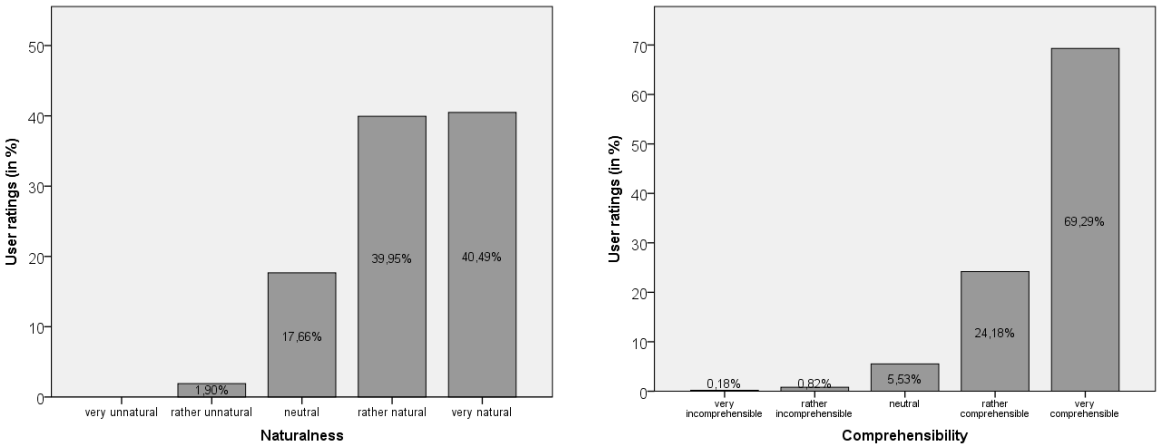
	SPDev	***DLDev***	***LPDev***
BL	0.809	5.628	0.167
VOS	1.029	13.241	0.236
MCV	1.063	11.949	0.190
RCV	1.069	13.218	0.187
BL vs. VOS	$Z =$ -8.344 *** ($r = .36$)	$Z =$ -18.661 *** ($r = .79$)	$Z =$ -15.693 *** ($r = .67$)
MCV vs. RCV	$Z =$ -0.731 ($r = .03$)	$Z =$ -6.116 *** ($r = .26$)	$Z =$ -0.543 ($r = .02$)

Note: *$p < .05$, **$p < .01$, ***$p < .001$
Effect size $r = .10$ (small effect), $r = .30$ (medium effect), $r = .50$ (large effect)

In a second step, driving performance was compared during voice output sequences in dependence of the sentence types MCV and RCV. While *SPDev* and *DLDev* indicated higher values for RCV (*SPDev* 1.069, *DLDev* 13.218) compared to MCV (*SPDev* 1.063, *DLDev* 11.949), the opposite was observed for *LPDev* (MCV 0.190, RCV 0.187). Wilcoxon signed-rank tests did not reveal clear differences in the case of *SPDev* and *LPDev*. In contrast, a significant degradation in terms of *DLDev* became apparent in dependence of the syntactic form of voice output ($Z =$ -6.116, $p < .001$, $r = .26$). Thus, a greater deviation in the distance to the lead vehicle was observed during voice output in the form of RCV compared to MCV. These findings are contrary to the Hypotheses 3.7 and 3.9, but support Hypothesis 3.8.

3.3.2.4 Subjective Assessment of User Experience

The participants of this user study were asked to repeatedly assess voice output of differing syntactic complexities. Overall, the syntactic paraphrases for MCV and RCV were assessed 368 times (8 per participant) and 1,104 times (24 per participant) concerning their perceived naturalness and comprehensibility, respectively. Figure 3.27 summarizes the results. Altogether, the voice prompts were rated as very good, within the higher values of the 5-point Likert scale, with $M = 4.19$ for naturalness ($Mdn = 4$, $SD = 0.78$) and an even higher mean



(a) Overall, the syntactic prompt paraphrases were assessed 368 times concerning their perceived naturalness with an average score of 4.19 ($Mdn = 4$, $SD = 0.78$).

(b) The syntactic prompt paraphrases were assessed 1,104 times concerning their perceived comprehensibility ($M = 4.62$, $Mdn = 5$, $SD = 0.65$).

Figure 3.27: Summary of user ratings. Adapted from Stier *et al.* (2020b, Figure 5), © 2021 Copyright held by the owner/author(s).

value of $M = 4.62$ for comprehensibility ($Mdn = 5$, $SD = 0.65$). Overall, the perceived comprehensibility was rated higher than the perceived naturalness of voice prompts (Wilcoxon signed-rank test, $Z =$ -4.554, $p < .001$, $r = .67$). The ratings regarding these variables strongly correlated with each other ($r = .532$, $p < .001$). Accordingly, if the comprehensibility increased, the naturalness increased analogously.

As noted above, a repeated-measures design was employed in this user study. For this reason, further exploratory evaluations concerning the perception of syntactic forms in voice output were conducted fitting two two-level GLMMs (`GENLINMIXED` procedure in `SPSS` v24.0; s. Appendix A.4.8.2). Similar to the evaluation presented in Section 3.2, subjects were introduced as random intercepts to account for the repeated-measures character of the collected data (*cf.* Heck *et al.*, 2013). Given the ordinal scale of the dependent variables $Naturalness$ and $Comprehensibility$, a cumulative logit link function was chosen. The results presented below will thus refer to the OR [95% CI] to represent the probability of an outcome to occur in a condition compared to the probability to occur in the absence of that condition. In the case of multiple comparisons, the sequential Bonferroni correction was applied. The parame-

Table 3.14: Fixed effects of the two two-level GLMMs for the dependent variables *Naturalness* and *Comprehensibility* (left), and their observed main effects (right).

	Parameter	**Levels**	*Naturalness*	*Comprehensibility*
	Complexity	AUT, MAN	n.s.	n.s.
	Domain	COP, DAS	n.s.	n.s.
	Question type	What, How, When	–	n.s.
	Sentence type	MCV, RCV	F(1,318)= 63,373 ***	F(1,1049)= 98,569 ***
	Age	18-29, 30-44, 45-59, 60-70	F(3,318)= 11,200 ***	F(3,1049)= 5,211 **
	Gender	female, male	F(1,318)= 71,890 ***	F(1,1049)= 14,731 ***
Exp[a]	*COP*	low < mid < high	n.s.	n.s.
Exp[a]	*DAS*	low < mid < high	n.s.	F(2,1049)= 3,248 *
Exp[a]	*Linguistics*	low < mid < high	F(2,318)= 17,284 ***	F(2,1049)= 8,847 ***
BFT[b]	*Openness*	mid < high	F(1,318)= 12,098 **	n.s.
BFT[b]	*Conscientiousness*	mid < high	F(1,318)= 5,409 *	n.s.
BFT[b]	*Extraversion*	mid < high	n.s.	n.s.
BFT[b]	*Agreeableness*	mid < high	F(1,318)= 11,076 **	n.s.
BFT[b]	*Neuroticism*	low < mid < high	F(2,318)= 7,158 **	n.s.
TA[c]	*Competence*	mid < high	n.s.	n.s.
TA[c]	*Neg. attitude*	mid < high	n.s.	F(1,1049)= 7,721 **
TA[c]	*Pos. attitude*	mid < high	n.s.	F(1,1049)= 9,214 **
TA[c]	*Enthusiasm*	low < mid < high	F(2,318)= 9,155 ***	F(2,1049)= 4,135 *

[a] Experience, [b] Big Five Traits, [c] Technical Affinity
Probability distribution: multinomial; link function: logit (cumulative).
Note: *$p < .05$, **$p < .01$, ***$p < .001$; n.s. not significant

ters and their respective levels listed in Table 3.14 (left) were included as fixed effects in both GLMMs (except for the parameter *Question type*, which was only available for the evaluation of *Comprehensibility*). As in Section 3.2, the values of metric and ordinal variables were recoded into subgroups for the purpose of better interpretability. An overview of the recoding procedure is provided in Appendix A.4.8.1.

3.3.2.4.1 Factors Influencing the Perception of Voice Prompts. The main effects of the two GLMMs for the two dependent variables $Comprehensibility$ and $Naturalness$ are summarized on the right side of Table 3.14. A number of parameters revealed their influence on the perception of voice output in general.

Sentence Types. Among the system-related parameters, a significant difference in the assessment of voice output was observed in particular for the $Sentence\ type$ concerning the perceived $Naturalness$ (F(1,318)= 63,373, $p < .001$) and $Comprehensibility$ (F(1,1049)= 98,569, $p < .001$). *Post hoc* analyses suggested that in the case of MCV there was a higher probability of a voice prompt being perceived as natural (OR 23.865 [0.166, 3426.499]) and comprehensible (OR 2.413 [0.95, 61.459]) compared to RCV. This finding generally confirms Hypotheses 3.10 and 3.11. However, the comparisons between the $Sentence\ types$ on an individual level failed to be significant ($Naturalness\ p = .210$, $Comprehensibility\ p = .594$), thus there was no clear difference in the perception of the individual $Sentence\ types$. In addition to the exploratory GLMM approach, Wilcoxon signed-rank tests were conducted to enable a focused examination of the $Sentence\ types$ in absence of further predictors. It became apparent that MCV was assessed as more natural ($M = 4.26$, $Mdn = 4$, $SD = 0.76$; Z = -3.135, $p = .002$, $r = .23$) and better comprehensible ($M = 4.69$, $Mdn = 5$, $SD = 0.58$; Z = -7.53, $p < .001$, $r = .32$) than RCV (Naturalness: $M = 4.12$, $Mdn = 4$, $SD = 0.81$; Comprehensibility: $M = 4.54$, $Mdn = 5$, $SD = 0.70$).

Age & Gender. In particular person-related parameters revealed an influence on the perception of voice output. The perceived $Naturalness$ and $Comprehensibility$ thus showed to be dependent on both the Age (F(3,318)= 11,200, $p < .001$; F(3,1049)= 5,211, $p < .001$) and $Gender$ (F(1,318)= 71,890, $p < .001$; F(1,1049)= 14,731, $p < .001$) of participants. While female subjects were in principle more likely to rate voice output as more comprehensible (OR 0.133 [0.050, 0.354]) and natural (OR 0.007 [0.002, 0.023]) than male subjects, this was especially true for the two older groups of participants with an age between 45 and 70 years compared to the 18-44 years old participants. Concering $Naturalness$ (F(3,318)= 11,200, $p < .001$), the odds ratio of 44-56 years old and 60-70 years old participants was 0.008 [0.001, 0.056] times and 0.003 [0.000, 0.052] times higher than that of 18-29 years old ($p < .001$), respectively. Similarly, their odds was 0.018 [0.003, 0.118] times and 0.007 [0.000, 0.107] times higher than that of participants between 30 and 44 years ($p < .001$). No significant difference was observed between 18-44 years and 45-70 years old participants. Concerning the perceived $Comprehensibility$ (F(3,1049)= 5,211, $p < .001$), a similar behav-

ior was observed with an odds ratio of 0.073 [0.018, 0.306] times and 0.113 [0.019, 0.659] times higher for 45-59 years old participants compared to participants aged between 18-29 ($p < .001$) and 30-44 ($p = .015$). Again, no clear difference was observed between 18-44 years and 45-70 years old subjects.

Prior Experience. Furthermore, significant differences in the perception of voice output became apparent for prior experiences. As for $Linguistics$ and $Naturalness$ ($F(2,318) = 17,284$, $p < .001$), highly experienced participants were more likely to assess voice prompts as natural than participants with a low (OR 1.195E-06 [2.888E-08, 4,948E-05], $p < .001$) or average (OR 0.032 [0.010, 0.097], $p < .001$) experience. In addition, the odds of participants with an indicated average linguistic knowledge was 0.004 [0.000, 0.001] times higher than for low experienced subjects ($p < .001$). A similar observation was made for the perceived $Comprehensibility$: The probability of participants with an indicated high linguistic knowledge to rate voice prompts as comprehensible was higher than for participants with an average (OR 0.251 [0.082, 0.764], $p = .015$) or low (OR 0.021 [0.003, 0.154], $p < .001$) linguistic experience. At the same time, the odds of ordinary experienced subjects was 0.082 [0.009, 0.753] times higher than for low experienced ones ($p = .027$). In the case of prior knowledge in the domain DAS, no differences were observed concerning the perceived $Naturalness$ of voice output. In contrast, experience in this domain revealed an influence on whether voice promtps were perceived as comprehensible ($F(2,1049) = 3,248$, $p = .039$). Here, the probability of low experienced participants to rate prompts as comprehensible was higher than for highly experienced subjects (OR 0.152 [0.025, 0.921], $p = .040$), while no clear difference was found between participants with an average experience and participants with more and less prior knowledge.

Big Five Traits. In the separate evaluation of voice output concerning its perceived $Naturalness$ and $Comprehensibility$, it became apparent in the analyses of this user study that the Big Five personality was particularly related to the former variable, while the Technical Affinity of participants was mainly related to the latter. No significant difference in the perceived $Comprehensibility$ of voice prompts was induced by the Big Five traits. Concerning $Conscientiousness$ ($F(1,318) = 5,409$, $p = .021$), the odds ratio to rate voice prompts as natural was 0.019 [0.001, 0.266] times higher for high manifestations of this trait compared to less conscientious participants. The opposite was observed for the traits $Agreeableness$ ($F(1,318) = 11,076$, $p < .001$), $Neuroticism$ ($F(2,318) = 7,158$, $p < .001$) and $Openness$ ($F(1,318) = 12,098$, $p < .001$). Here, the probability of highly agreeable and open partici-

pants to assess voice prompts as natural was lower than for participants assigned to a lower level of these traits ($Agreeableness$: OR 8.370 [1.056, 66.312]; $Openness$: OR 4.812 [1.466, 15.795]). Similarly, average neurotic participants were more likely to assess voice output as natural compared to highly neurotic subjects (OR 0.073 [0.006, 0.835], $p = .035$). At the same time, no clear difference was observed between participants assigned to the subgroups high and low or average and low neurotic.

Technical Affinity. With respect to the Technical Affinity component $Enthusiasm$ (F(2,318)= 9,155, $p < .001$), the odds of low enthusiasts to rate voice prompts as natural was 102.554 [8.098, 1298,718] times lower compared to average enthusiastic ($p < .001$) and 103.382 [5.101, 2095.424] times lower compared to highly enthusiastic participants. No clear difference was found in the comparison between the participants which indicated an average or high manifestation of this trait. A similar behaviour was found for the Technical Affinity components $Negative\ Attitude$ (F(1,1049)= 7,721, $p = .006$) and $Positive\ Attitude$ (F(1,1049)= 9,214, $p = .002$) concerning the perceived $Comprehensibility$ of voice output, where high manifestations of these traits indicated a higher probability to assess prompts as comprehensible compared to less negatively (OR 0.058 [0.009, 0.366]) and less positively (OR 0.249 [0.089, 0.701]) positioned participants with respect to technical devices. Concerning $Enthusiasm$ (F(2,1049)= 4,135, $p = .016$), the odds of average enthusiastic participants was 0.060 [0.011, 0.328] times higher than for low enthusiasts ($p < .001$). Simultaneously, no clear difference was observable between highly enthusiastic participants and lower manifestations of this trait.

3.3.2.4.2 Factors Related with the Syntactic Complexity of Voice Prompts on their Perception. In addition to the main effects described above, interaction effects between the fixed effects displayed in Table 3.14 and the parameter $Sentence\ type$ were investigated in order to specify the relationship between them and the perception of syntactically differing voice prompts. The results are summarized in Table 3.15. In addition, an overview of the conducted *post hoc* analyses in the case of significant interaction effects is provided in Table 3.16.

Domain & Question Type. In this context, both the $Domain$ (F(1,1049)= 5.032, $p = .025$) and $Question\ type$ (F(2,1049)= 3.466, $p = .032$) appeared to be related with the perceived $Comprehensibility$ of syntactically different voice prompts. *Post hoc* analyses revealed that in the case of DAS there was a higher probability to rate MCV as comprehensible compared to COP (OR 0.512 [0.285, 0,920]). Consequently, in the case of COP, participants were less likely

Table 3.15: Interaction effects with the parameter *Sentence type*. Adapted from Stier *et al.* (2020b, Table 2), © 2021 Copyright held by the owner/author(s).

	Parameter	**Levels**	*Naturalness*	*Comprehensibility*
	Complexity	AUT, MAN	n.s.	n.s.
	Domain	COP, DAS	n.s.	$F(1,1049) = 5.032$ *
	Question type	What, How, When	–	$F(2,1049) = 3.466$ *
	Age	18-29, 30-44, 45-59, 60-70	n.s.	n.s.
	Gender	female, male	n.s.	n.s.
Exp[a]	*COP*	low < mid < high	n.s.	n.s.
Exp[a]	*DAS*	low < mid < high	n.s.	n.s.
Exp[a]	*Linguistics*	low < mid < high	$F(2,318) = 93.287$ ***	$F(2,1049) = 327.406$ ***
BFT[b]	*Openness*	mid < high	n.s.	n.s.
BFT[b]	*Conscientiousness*	mid < high	$F(1,318) = 5.093$ *	n.s.
BFT[b]	*Extraversion*	mid < high	n.s.	n.s.
BFT[b]	*Agreeableness*	mid < high	$F(1,318) = 6.300$ *	$F(1,1049) = 28.130$ ***
BFT[b]	*Neuroticism*	low < mid < high	$F(2,318) = 3.781$ *	n.s.
TA[c]	*Competence*	mid < high	n.s.	n.s.
TA[c]	*Neg. attitude*	mid < high	n.s.	n.s.
TA[c]	*Pos. attitude*	mid < high	n.s.	n.s.
TA[c]	*Enthusiasm*	low < mid < high	n.s.	n.s.

[a] Experience, [b] Big Five Traits, [c] Technical Affinity
Probability distribution: multinomial; link function: logit (cumulative).
Note: *$p < .05$, **$p < .01$, ***$p < .001$; n.s. not significant

to assess MCV as comprehensible than RCV. Furthermore, the odds ratio of voice prompts being rated as comprehensible as answers to the question When was 2.311 [1.186, 4.500] times lower for MCV and thus higher for RCV compared to What ($p = .014$). No significant difference was found in the comparison between How and When concerning the preference for a *Sentence type*, however, the results based on the odds ratio suggested a higher probability that participants rated MCV as comprehensible in the case of How compared to When (OR 1.666 [0.907, 3.058]). Similarly, an insignificant trend indicated that MCV was less likely perceived as comprehensible in an answer to the question How compared to What (OR 0.721

[0.341, 1.524]).

Prior Experience. The linguistic knowledge of participants was observed to be related with both the perceived $Naturalness$ (F(2, 318)=93.287, $p < .001$) and $Comprehensibility$ (F(2, 1049)= 327.406, $p < .001$) of voice prompts in dependence of their syntactic complexity. *Post hoc* comparisons revealed a higher probability of MCV being rated as natural by participants with an indicated low linguistic knowledge compared to average experienced (OR 3.19e9 [1.26e8, 8.08e10], $p < .001$) or highly experienced (OR 4.80e8 [1.93e8, 1.19e11], $p < .001$) subjects. Similarly, MCV was more likely being perceived as comprehensible for participants with a low level in $Linguistics$ compared to participants with an average level (OR 5.84e5 [1.14e5, 2.99e5], $p < .001$) or a high level (OR 1.44e6 [4.71e5, 4.42e5], $p < .001$). No significant differences became apparent in the comparisons between the parameter levels high and mid, however, trends suggested that the odds ratio for MCV being rated as natural was 1.501 [0.687, 3.281] times and being rated as comprehensible 2.470 [0.891, 6.846] times higher for particpants with an average prior knowledge in $Linguistics$ compared to highly experienced ones.

Big Five Traits. While the Technical Affinity was not found to be related with the perception of syntactically differing voice output, this was the case for Big Five personality traits. In this context, the traits $Conscientiousness$ (F(1,318)= 5.093, $p = .025$), $Agreeableness$ (F(1,318)= 6.300, $p = .013$) and $Neuroticism$ (F(2,318)= 3.781, $p = .024$) indicated differerences in the acceptance of syntactically differing voice prompts among their respective manifestation levels. As such, participants assigned to the level mid were more likely to rate MCV as natural ($Conscientiousness$: OR 9.631 [1.337, 69.387]; $Agreeableness$: OR 15.635 [1.812, 134.929]) and comprehensible ($Agreeableness$: OR 7.468 [3.549, 15.712]) than participants with a higher manifestation of these traits. Consequently, in the case of highly conscientious or agreeable participants, a higher probability was observed to assess RCV as natural and comprehensible compared to MCV. As for the trait $Neuroticism$, the odds ratio for highly neurotic participants to perceive MCV as natural was 28.642 [2.283, 359.338] times and 19.797 [2.168, 180.752] times lower than for participants who indicated a low ($p = .009$) or average ($p = .008$) manifestation of this trait. The comparison between the lower levels of $Neuroticism$ did not reveal a significant difference concerning the preference of a $Sentence\ type$, but the results suggested the trend that low neurotics were more likely to assess MCV as natural than average neurotic participants (OR 1.447 [0.358, 5.843]).

Table 3.16: *Post hoc* analyses for significant interaction effects with the parameter *Sentence type* and syntactic preferences. Adapted from Stier *et al.* (2020b, Table 3), © 2021 Copyright held by the owner/author(s).

Parameter levels		***Naturalness*** Odds ratio [95% CI]		***Comprehensibility*** Odds ratio [95% CI]		**Interpretation:** MCV ⟵ preference ⟶ RCV
Domain	DAS vs. COP	n.s.		0.512	[0.285, 0.920] *	DAS – – – COP
Q.type	When vs. What	–		2.311	[1.186, 4.500] *	What – How – When
	When vs. How			1.666	[0.907, 3.058]	
	What vs. How			0.721	[0.341, 1.524]	
Linguist.	high vs. low	4.80e9	[1.93e8, 1.19e11] ***	1.44e6	[4.71e5, 4.42e5] ***	low – mid – high
	mid vs. low	3.19e+9	[1.26e8, 808e10] ***	5.84e5	[1.14e5, 2.99e5] ***	
	high vs. mid	1.501	[0.687, 3.281]	2.470	[0.891, 6.846]	
Consc.	high vs. mid	9.631	[1.337, 69.387] *	*n.s.*		mid – – – high
Agree.	high vs. mid	15.635	[1.812, 134.929] *	7.468	[3.549, 15.712] ***	mid – – – high
Neurot.	high vs. low	28.642	[2.283, 359.338] **	*n.s.*		low – mid – high
	mid vs. low	1.447	[0.358, 5.843]			
	high vs. mid	19.797	[2.168, 180.752] **			

Probability distribution: multinomial; link function: logit (cumulative).
Comparisons are based on the first named parameter and RCV as referent.
Note: *$p < .05$, **$p < .01$, ***$p < .001$

3.3.3 Discussion of Results

This section discusses the results of the present user study and reflects on the hypotheses formulated in Subsection 3.3.1.6. A summary of the hypotheses and their validation is provided in Table 3.17. Subsequently, the results of the exploratory evaluation concerning the subjective assessment of user experience will be summarized.

- The results show that the two driving conditions AUT and MAN induce a different degree of cognitive load on the participants in the driving simulator. The individual DALI dimensions concerning the effort of attention, stress, interference and visual, temporal and auditive demands indicate a significantly higher cognitive load for the manual driving MAN compared to the autonomous driving part AUT. Thus, with an increasing degree of automation, the cognitive load decreases. This finding supports Hypothesis 3.3.
- The driving performance measures *SPDev*, *DLDev* and *LPDev* were significantly higher during voice output sequences compared to the baselines without any voice-based interaction. These objective measures thus indicate a degradation of the participants' driving performance with parallel voice output. Therefore, the results coincide with the general consent that driving performance is influenced by the secondary task (limited to voice output sequences, in the present case) given an increased cognitive load. These findings confirm Hypotheses 3.4, 3.5 and 3.6.
- The examination of driving parameters further shows a direct influence of syntactic forms and their inherent complexity on driving performance. An increased deviation in the distance to the lead vehicle for RCV and the associated degradation of driving performance indicates that syntactic complexity is reflected in an increased cognitive load. This finding supports Hypothesis 3.8. Contrary to the expectations, this behavior is not found for *SPDev* nor *LPDev*. Like *DLDev*, *SPDev* exhibits increased values for RCV compared to MCV, while this effect is reversed for *LPDev*. Although research exists investigating that at increased cognitive load a microsteering behavior occurs and improves lane keeping (*e.g.*, Li *et al.*, 2018), the same relation for the initial baseline comparison would then have been expected. Due to the lack of a clear result at this point with the available performance measures, the Hypotheses 3.7 and 3.9 are therefore rejected. However, the choice of a particular syntactic structure had a perceivable influence on the performance of the primary driving task. This observation indicates that the design of vehicle-related voice output should be carefully chosen to minimize

safety-critical aspects, such as cognitive load and driver distraction.

- The results show that the perception of voice prompts in terms of their perceived naturalness and comprehensibility depends on their syntactic forms. More precisely, the analyses suggest the trend of a higher probability for the syntactically simpler sentence type MCV being perceived as natural and comprehensible while driving compared to the syntactically complex sentence type RCV. Although this direct comparison of sentence types failed to be significant in the employed exploratory GLMM approach, additional Wilxocon signed-rank tests conducted to enable a focused examination of MCV vs. RCV in absence of further model predictors confirmed a clear preference for voice prompts in the form of MCV over RCV in terms of their perceived naturalness and comprehensibility. This finding thus supports Hypotheses 3.10 and 3.11. Due to a strong positive correlation of the perceived naturalness and comprehensibility of voice output, a possible overlap of these concepts was observed in the subjective perception of the participants. With an increase in the perceived comprehensibility, the perceived naturalness increases analogously. Thus, the separate assessment of the concepts naturalness and comprehensibility does not seem required in the context of in-vehicle voice output.

In addition to the validation of hypotheses, an exploratory evaluation was conducted in order to identify individual user and system parameters, which influence the perception of voice prompts in general. Besides the influence of the syntactic form of voice output on its perceived naturalness and comprehensibility, in particular person-related characteristics turn out to be further influencing factors. For instance, women and participants in the older age groups between 45 and 70 years generally rate voice output as more comprehensible and natural. Male and younger participants between 18 and 44 years appear comparatively more critical in their evaluations. In addition, a relationship between the perception of voice output and the degree of prior experiences can be observed. With an increasing linguistic expertise, voice output is perceived as both more natural and comprehensible. In this context, it is reasonable to assume that with a linguistic affinity, speech can be processed

Table 3.17: Validation of Hypotheses.

$DALI_{AUT} < DALI_{MAN}$	✓	(3.3)
$SPDev_{BL} < SPDev_{VOS}$	✓	(3.4)
$DLDev_{BL} < DLDev_{VOS}$	✓	(3.5)
$LPDev_{BL} < LPDev_{VOS}$	✓	(3.6)
$SPDev_{MCV} < SPDev_{RCV}$	(✗)	(3.7)
$DLDev_{MCV} < DLDev_{RCV}$	✓	(3.8)
$LPDev_{MCV} < LPDev_{RCV}$	(✗)	(3.9)
$Nat_{RCV} < Nat_{MCV}$	✓	(3.10)
$Comp_{RCV} < Comp_{MCV}$	✓	(3.11)

more routinely and intuitively, and thus even SDS voice output that has never been heard before is perceived as natural and comprehensible. This explains the lower probability for less experienced participants to rate voice output as natural and comprehensible, since they have comparatively less routine in handling speech. The opposite is true for domain-specific knowledge. Here, the results for the domain DAS show a decrease in the perceived naturalness and comprehensibility with an increasing level of experience. At first, this observation seems counterintuitive, since a particular topic is generally better understood with an appropriate prior knowledge. Given that DAS-experienced subjects received the same explanatory voice prompts as the less experienced ones and that the perceived naturalness correlates strongly with comprehensibility, the evaluation behavior of experienced participants seems to be explained less by the structural form of voice output but rather on a content-related level. It is reasonable to assume that domain-experienced subjects relate the content of voice prompts to their prior experience and own knowledge. Since the voice prompts used in this study provided basic explanations of various vehicle functions and were not created for the purpose of technical immersion, they possibly appeared too general and superficial to experienced participants, which is reflected in their subjective ratings. Although the content of voice output was not the focus of this study, the findings described suggest that this level should also be considered for the goal of a satisfactory user experience. Due to the focus on syntactic forms in voice output, this content-related aspect is outside the scope of this work and is referred to as an object of investigation of future related research.

A relationship between the perception of voice output and the personality traits of SDS users has already been discussed in Section 3.2.3. In addition to the fact that a person's personality is directly reflected in his or her linguistic behavior, the results of the present study likewise demonstrate that individual Big Five traits are related with the perceived comprehensibility and naturalness of voice output. The more conscientious a person is, the more likely he or she is to perceive voice output as natural. The opposite is indicated for the traits Agreeableness, Neuroticism and Openness. Here, a decrease in the probability to rate voice prompts as natural is observed with an increased manifestation of these traits. With a focus on personal attitudes regarding technical devices, the perception of voice output is also related to a person's technical affinity. Thus, voice prompts are assessed as more comprehensible and natural with an increasing degree of enthusiasm and a positive as well as critical attitude.

Overall, the perception of voice output is generally dependent on several factors that should be considered in the design of voice prompts for the most positive user experience. In order to further specify the influence of syntactic forms with regard to the present study, interaction

effects were considered.

The investigation of the direct relationship between sentence types with a differing syntactic complexity and the user and system parameters allowed a more concrete identification of relevant factors to be considered in the design of voice output. The results show that both the domain and question type represent dependent factors in the perceived comprehensibility of syntactically differing voice prompts. While a higher probability was found for RCV being rated as comprehensible in the context of COP and answers to the question When, this holds for MCV in the context of DAS and explanatory prompts for the questions What and How. Similar to the observation concerning prior experiences above, this behavior should be interpreted under consideration of a content-related level as the distinguishing factor. While COP exclusively provides explanations concerning the recurring ensemble playing of various in-vehicle programs, such as music, fragrance, lighting and massage, which focus on the well-being of a driver, the contents for DAS are broader and more varied, ranging from an explanation for the cause, action and purpose of a driving assistant. Similarly, voice prompts as answers to the questions What and How supply varied explanations as opposed to When, which as a special case of How (s. Subsection 3.3.1.3.3) specifically refers to system limitations. The comparably complex and varied contents of DAS, What and How are thus reflected by a preference for simpler, linear structures. As soon as the content gets less varied and simpler, RCV is the preferred sentence type. Individual user characteristics were furthermore identified as dependent factors in the perception of syntactically differing promtps. Overall, the results indicate an increased probability for RCV being rated as natural and comprehensible the stronger a participant is associated to a personality trait (Conscientiousness, Agreeableness, Neuroticism) or experienced in Linguistics. The less prevalent these characteristics are, the more likely the preference for MCV increases. Concerning the linguistic experience, this observation seems to contradict the results of Pilot Study 2 (s. Section 3.1.6), where linguistically experienced participants (at the latest in Part 3, s. Figure 3.4) were aware of linguistic cues, had a clear opinion regarding the appropriateness of syntactic structures and preferred MCV in voice output. However, since syntactic forms were not revealed as an object of investigation in the present study, the intuitiveness of user assessments is assumed. Accordingly, the contradictory results seem to be related with the additional secondary driving task in this study. Additionally, as outlined before, it is assumed that linguistically affine participants are used to process syntactically complex structures and thus evaluate them as more natural and comprehensible. In contrast, the connection between the preference for syntactic structures and the personality traits of a person is not as easy to capture. Why conscientious persons

show an affinity for syntactically more complex structures than less conscientious persons is not directly derivable from this personality trait. However, as introduced in Section 3.2.3, this question can be extended with respect to a person's own linguistic behavior by investigating whether he or she prefers an SDS with similar or opposite personality (*attraction* vs. *complementarity principle*) and, consequently, similar or different linguistic behavior. To the best of our knowledge, no prior work exists examining the relationship between syntactic structures and user personality in an automotive context with SDS interaction as a secondary task. For this reason, further research will be required at this point.

It should be mentioned that no relation was observed between the driving task's complexity (AUT, MAN) and sentence types. Thus, considering user preferences, the long-term goal to design SDS interaction according to interpersonal models seems independent of the SAE level.

3.4 Summary of Results

This section provides a summary of the conducted user studies on language perception. First, the approach to manually prepare syntactic paraphrases is described briefly, before the two driving simulator studies and their main findings are summarized. Finally, implications on the following research are outlined.

3.4.1 Summary

In this section, two studies on the perception of syntactic forms in in-vehicle voice output were presented. For this purpose, paraphrases with different syntactic complexity were manually created for the context of conceptual explanations embedded in one-shot QAS. Since no prior work was available as a basis, a custom approach was developed to ensure a comparable complexity of the explanations independent from their content, and thus allowing valid inferences about the syntactic differences of the created paraphrases (s. Section 3.1). Based on the assignment of instruction manual contents for different driving assistants to components answering individual question types, the presented approach relied on a general semantic-syntactic structure for conceptual explanations according to the model of a frame in FrameNet (Baker *et al.*, 1998). By the application of several surface measures, the semantic complexity

was additionally ensured. The base texts produced by this method were syntactically paraphrased with the realization of various aggregation strategies. The syntactic paraphrases created in this way were synthesized and served as voice prompts for further investigations. The general validity of the generation approach was examined in a first pilot study (s. Section 3.1.5). Here, all participants confirmed a semantic and structural comparability across vehicle functions and syntactic features. Additionally, a second pilot study investigated whether the approach of collecting subjective user preferences regarding syntactic forms in voice output is a valid elicitation method (s. Section 3.1.6). The results indicated that although syntactic differences were generally not directly perceived and named via audio, subconscious differentiation and prioritization nevertheless occurred, reflected in a majority preference for simple syntactic forms. Thus, the structural complexity associated with different syntactic forms has been shown to be a distinguishing feature for the perception of voice output. Based on this, it was argued that the lack of awareness for syntactic differences allows for intuitive, unbiased user preferences and that the evaluation of voice output with respect to individual syntactic preferences represents a valid methodology. Finally, the presented approach was applied to comfort functions. Although it was directly translatable to this second domain, the different focus of the domains revealed only limited comparability of the resulting conceptual explanations in terms of their semantic-syntactic complexity. The extent to which this difference, inherent in the domains, affects the perception of syntactic forms in voice output should be investigated in the following user studies.

An initial driving simulation study was conducted to investigate influencing factors on the perception of voice output and the role of syntactic forms while driving (s. Section 3.2). For this purpose, the participants interacted with a simulated SDS in two driving conditions and evaluated its voice output within the two domains DAS and COP. Overall, the results of this study indicated that both user and system characteristics influence the perception of voice output. For instance, it was demonstrated that the perception of voice output and its syntactic complexity depend on individual prior experiences and personality traits of a user. However, against the background of the high cognitive load of the participants induced by the study design, which was subsequently classified as highly demanding, the results were interpreted as a basis for a further user study.

The study design of this second driving simulation study (s. Section 3.3) was optimized according to the weaknesses of the first study. This included the verification of fundamental assumptions: As intended, the study results showed a significantly higher cognitive load induced

by the manual driving section compared to autonomous driving. Furthermore, the results of objective driving parameters showed a degradation in driving performance while participants were listening to voice output compared to driving without voice interaction. These findings are consistent with previous research. Additionally, for voice prompts with a high syntactic complexity, a degradation in driving performance was observed in terms of the deviation in the distance to the lead vehicle. Contrary to expectations, the performance measures lane keeping and speed did not show an equally clear deviation in driving behavior. Besides the examination of driving performance measures, a further goal of this study was again to specifically identify factors related to the perception of voice output and to specify the influence of syntactic forms in voice output. Similar to the prior study, the results showed a direct influence of several user and system parameters on the perceived naturalness and comprehensibility of voice output. Against the background of this work, the influence of syntactic forms was of particular interest. Here, a general preference for syntactically simple SDS voice output was demonstrated as opposed to more complex syntactic structures. This coincides with the observations in the pilot studies. Additionally, the perception of voice output showed a dependency between its syntactic complexity and, for example, the domain, a person's personality and prior experiences. The more prominent linguistic knowledge and personality traits were, such as Conscientiousness, Agreeableness and Neuroticism, the more syntactically complex voice prompts were preferred. Explanations of DAS functions were considered more comprehensible compared to COP in the form of simple syntactic structures. Although the findings regarding individual parameters from the first study were not entirely confirmed, there are apparent overlaps between the two driving simulator studies overall. The differences in the results are attributed to the differences in their study designs. Thus, by reducing the complexity of the driving task in the second study, the results of the first study were revised and partially extended on a valid basis.

Overall, the research work presented in this section establishes the foundation for an adaptive strategy concerning the syntactic design of voice output in SDS interaction as a secondary task. In comparison to previous theory-oriented approaches, here, the focus was on experience and usability from a user perspective. The described results indicate that there is no rigid default solution for natural and comprehensible SDS voice output in dual-task environments. Rather, the appropriateness of and preference for a syntactic structure and its associated complexity depend on individual characteristics of the driver and the application context.

3.4.2 Implications on Research Work

In general, the assumptions made in this section were confirmed. However, the driving performance parameters available in this work did not show consistently clear results. At this point, further research would be required to measure the influence of syntactically different voice prompts on driving performance, which is beyond the scope of this work.

With respect to the subjective evaluations of the two user studies presented in this section, despite general overlaps also differences can be observed. As noted above, these differences are attributed to the highly demanding driving task of the first study. Thus, by reducing the complexity of the driving task, the results of the first study were revised and substantively extended. Therefore, in the remainder of this thesis, the results of the second driving simulation study will be used as a reference.

Overall, several system and user parameters revealed to be related with the perceived naturalness and comprehensibility of voice output. In the context of this work, particularly the influence of different syntactic forms was identified as a relevant aspect to be considered in the design of voice prompts. Towards the goal of intuitive, natural voice output following the human model, thus, the syntactic complexity of voice prompts should be adapted according to individual user and system parameters. However, it is beyond the scope of this work to account for all identified parameters. In this context, the relationship between the perception of voice output and a user's Big Five personality traits represent a particularly valuable framework to account for syntactic differences from a linguistic perspective. Given that human personality is directly reflected in language behavior, it represents an ideal basis to relate the here identified preferences concerning SDS voice output and syntactic forms with actual linguistic behavior. In this way, the question can be approached for the automotive context and SDS interaction as a secondary task whether a user with a particular personality prefers a similar or complementary SDS personality (*cf. attraction* vs. *complementarity* principle) with a respective linguistic behavior. Departing from the user studies on language perception, the following chapter will therefore consider the language production side and the linguistic behavior of SDS users in a dual-task environment against the background of their Big Five personality traits. By comparing the role of syntactic structures in human linguistic behavior with the indicated preferences of this chapter, it will be possible to elaborate a user-focused adaptation strategy while simultaneously taking the interaction context as a parallel secondary task into account.

Chapter 4

User Studies on Spoken Language Production

Human interlocutors adapt their language style to each other in interpersonal communication to interact efficiently (Pickering and Garrod, 2004). The principle of *alignment* is generally found at all linguistic levels, from lexis to syntax. For example, as demonstrated by Levelt and Kelter (1982) the question “What time does your shop close?” may be answered with a simple “5 o’clock”. If in contrast a prepositional phrase is included in the question (“At what time...”), the interviewer tends to adapt to this structure and is likely to include the preposition in his or her answer (“At 5 o’clock”), too.

In order to enable the most efficient form of interaction in the context of SDSs based on the human model, computers – in place of a human interlocutor – are expected to flexibly react according to individual requirements of a user, such as in the form of adaptive voice output that adapts to the user’s language style. In this context, it is common consent that more intelligent software is required to enable complex HMI (Jokinen, 2003) and that a computers’ responses need to be more sophisticated, while it is increasingly capable to understand what an SDS user says (Rambow *et al.*, 2001). This becomes even more relevant in dual-task environments like driving a car, where language interaction represents a secondary task in parallel to a prioritized primary driving task. The alignment of voice output to the linguistic behavior of a user represents a valuable contribution in this context: If the speech-based interaction with an SDS as a parallel secondary task can be designed as natural and intuitive as possible by means of linguistic alignment, a reduction of driver distraction and an increased safety are expected. For this purpose, individual requirements of an SDS user need to be

considered with respect to the driving situation as interaction context (*e.g.*, driving on a free highway vs. urban traffic with traffic light control). Furthermore, it is essential to understand the properties of a driver's language while driving in order to adequately design SDS voice output to be generated in such scenarios.

Against this background, this chapter focuses on the aspect of language production in the dual-task environment of driving from a driver's perspective. Towards the goal of a user- and situation-adaptive strategy for the syntactic design of in-vehicle SDS voice output, the impacts of human personality (as user characteristic) and the driving situation (as contextual characteristic) on human language production are investigated to characterize speech and to identify syntactic features, which are dependent upon the interaction context of driving. For this purpose, a large-scale driving simulation study was conducted in order to construct a corpus of spoken driver language as a basis for feature extraction and linguistic analyses with a focus on syntactic forms and complexity.

In this chapter, first, the data collection study is presented in Section 4.1 including a description of the employed methodology and the resulting spoken language corpus. Second, Section 4.2 provides the results of an exploratory Principal Component Analysis (PCA) for the purpose of better interpretability of the syntactic features of spoken language. On this basis, Section 4.3 describes the procedure and results of the linguistic analyses under consideration of the driving condition and the speaker's personality traits. Finally, a summary of the obtained observations is provided in Section 4.4.

4.1 Data Collection Study

In order to examine the linguistic behavior of drivers in spoken interaction as a secondary task, a driving simulation study was conducted to collect spoken language data. The data collection took place as a WoZ experiment, where the participants were asked to answer the small talk questions by a simulated voice assistant while driving (QAS with reversed conversation roles as defined in Section 3.1.1).

The following sections are based on the publications by Stier *et al.* (2020c) and Stier *et al.* (2020e). They present the methodology employed in the data collection study, followed by a description of the data processing and the resulting spoken language corpus.

4.1.1 Methodology

In this section, the methodological approach of the data collection study is outlined. First, the participants and the experimental design are described. Second, the data preparation procedure is explained as a basis for the analysis of the resulting spoken language corpus.

4.1.1.1 Participants

A total of 72 German native speakers between 22 and 66 years ($M = 40.00$, $SD = 13.28$) and a gender distribution of 46 male and 26 female subjects participated in the experiment. All of them possessed a valid driver's license and received an expense allowance of 50€ for participating.

4.1.1.2 Experimental Design

The study design was chosen against the background of collecting spoken language data while driving. For this purpose, different driving conditions were integrated into the within study design. As shown in Figure 4.1, the linguistic behavior of participants was obtained under three driving conditions (parked position, highway, city). During the first part, the participants were supposed to get used to the situation of answering small talk questions from a simulated voice assistant while parked. The driving sections on a highway and in a city were then intended to collect data on the language behavior of the participants in driving situations with an increasing level of complexity.

4.1.1.3 Materials

This section provides an overview of the materials used in this data collection study.

4.1.1.3.1 Questionnaires. In this study, two questionnaires were employed to capture demographic and personal data from the participants. They were created using the tool `soSci` and are found in Appendix B.1.

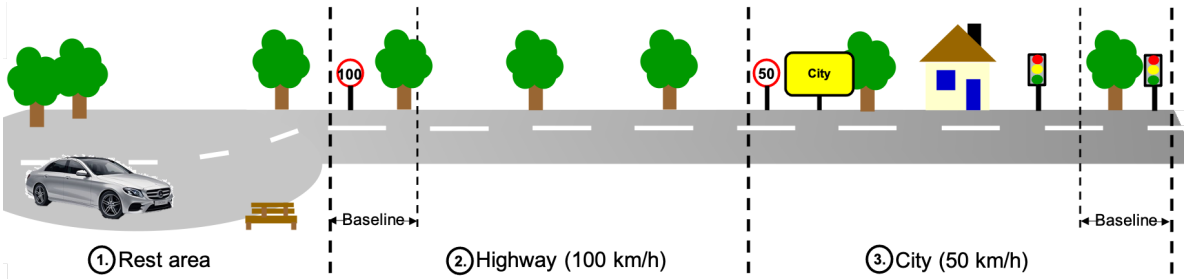


Figure 4.1: Spoken language was collected under three different driving conditions (parked position, highway, city). Taken from Stier *et al.* (2020c, Figure 2) with kind permission from Association for Computing Machinery.

- **Preliminary Questionnaire:** Demographic information (age, gender, etc.) about the participants was collected in a pre-survey. They were furthermore asked to estimate their level of lingiustic talent on a 5-point Likert scale.
- **Big Five Personality Traits:** The participants were additionally asked to self-assess their personality traits by means of the German version of the BFI questionnaire according to Rammstedt and Danner (2016) on a 5-point scale. It consists of 45 questions assigned to the five personality traits Agreeableness, Conscientiousness, Extraversion, Neuroticism and Openness. It is a frequently applied instrument to reduce human behavior to a small number of interpretable dimensions (s. Section 2.4).

4.1.1.3.2 Small Talk in Question-Answer Sequences. In order to be able to collect speech data from the participants, the previously assigned conversational roles were reversed within QAS. Thus, in this study, the simulated voice assistant asked questions to be answered by the participants. For this purpose, small talk was chosen as interaction topic in order to record the individual language use of participants independently of common voice-controlled vehicle-related functions. All questions were based on simple, personal experiences and private preferences in order to allow the participants to easily answer them with reference to their own daily environment. Depending on the length of a particpant's answers, the voice assistant was able to ask more or less questions from four small talk topics (general small talk, leisure time, preferences, travelling; s. Table 4.1). A complete list of all small talk questions used in this study is provided in Appendix B.2.

Table 4.1: Small talk topics and example questions. Adapted from Stier *et al.* (2020c, Table 1) with kind permission from Association for Computing Machinery.

Topic	**Number**	**Example**
General small talk	14	*Was würdest Du tun, wenn Du morgen im Lotto gewinnen würdest? Welche Wünsche würdest Du Dir gerne erfüllen?* (eng. "What would you do if you won the lottery tomorrow? Which wishes would you like to fulfill?")
Leisure time	11	*Was ist dein Lieblingshobby und wie kam es zu diesem Hobby?* (eng. "What is your favourite hobby and how did this hobby come about?")
Preferences	12	*Mein Lieblingsessen ist Lasagne. Was hälst du von Lasagne? Was ist dein Lieblingsessen?* (eng. "My favorite food is lasagna. What do you think of lasagna? What's your favorite food?")
Travelling	13	*Wie findest Du Kreuzfahrten? Welche Erfahrungen hast Du damit bisher gemacht?* (eng. "What do you think about cruises? What experiences have you made with them so far?")

Figure 4.2: Participants chose between Petra (left) and Yannick (right) as conversational partner. Taken from Stier *et al.* (2020c, Figure 1) with kind permission from Association for Computing Machinery.

4.1.1.3.3 An Interlocutor in the Form of an Avatar. In order to establish a more personal relationship between the driving participant and the simulated voice assistant, each participant was asked to select an avatar prior to the start of the study. It was then displayed in the HU screen during the entire journey (s. Figure 4.2). The introduction of a visually recognizable interlocutor was intended on the one hand to strengthen trust in the voice assistant and thus on the other hand to increase the participants' willingness to answer the small talk questions posed by the voice assistant.

4.1.1.3.4 Wizard-of-Oz as Simulated Spoken Dialog System. A Daimler internal tool on the model of SUEDE (Klemmer *et al.*, 2000) was used to simulate spoken interaction between the participants and a real SDS. For this purpose, the dialog flow of the previous user studies on language perception (s. Sections 3.2 and 3.3) was simplified as schematically visualized in Figure 4.3.

After the experimenter (green) started the dialog under consideration of the participant's avatar selection, the avatar was projected as a picture on the HU screen. In a next step, the experimenter initiated a dialog task in the form of a WoZ question via a corresponding button in the WoZ tool. The questions by the WoZ were synthesized in advance using TTS synthesis

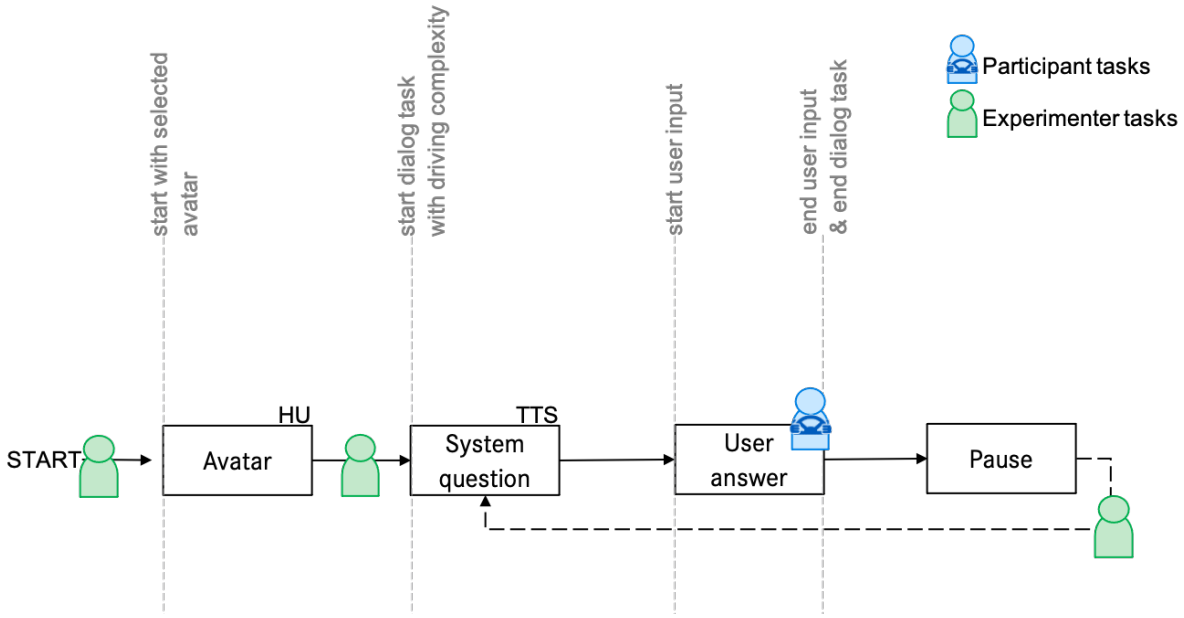


Figure 4.3: Schematic representation of the dialog flow using the WoZ tool.

(Nuance Vocalizer Studio 3.0.2[1], female voice: Petra-ML, or male voice: Yannick-ML, according to the chosen avatar) and integrated into the tool as WAV files. After the participant's answer (blue) and a pause of six seconds in order not to expose the participants to time pressure, the experimenter could initiate the next dialog task. This procedure was repeated until the specified driving time ended. The information about the driving condition and dialog flow (grey markers) were logged by the WoZ tool.

4.1.1.3.5 Experimental Setup. The study was conducted in a fixed-base driving simulator of a Mercedes-Benz E-Class at the Daimler site in Sindelfingen, Germany. For this purpose, the test environment of the user study in Section 3.3 (p. 91) was adopted by only exchanging the dialog tasks displayed on the HU screen by the participant's selected avatar. The voice interaction of the participants with the simulated voice assistant was recorded using cameras. In addition, the driving behavior was measured via RTDD, which were synchronized with the Wizard tool via a CAN bus system.

[1] `https://www.nuance.com/de-de/omni-channel-customer-engagement/voice-and-ivr/text-to-speech/vocalizer.html` (Online 12/09/2020)

4.1.1.4 Procedure

The procedure of this study was divided into two parts and lasted approximately 60 minutes per particpant. The following subsections provide a detailed overview of the individual phases.

4.1.1.4.1 Phase 1: Pre-Survey and Instructions. Prior to the start of the experiment, each participant was asked to sign a declaration of consent to the collection of personal data and recording of sound material, as well as a non-disclosure agreement. Subsequently, the participants completed the pre-survey questionnaire (s. Appendix B.1) and received an introduction into the study content and procedure (s. Appendix B.3). They were instructed to interact with a voice assistant during a daytime drive on different route sections. After an initial part parked on a rest area, the participants were asked to maintain a constant speed of 100 km/h on a highway and 50 km/h in a city. They were additionally instructed to follow a lead vehicle at a distance of approximately 100 m (*i.e.*, two delineator posts). The participants were prepared for the spoken interaction with the voice assistant in the form of a small talk conversation covering several topics initiated by the voice assistants' questions. They were assured that there were no wrong or right answers, but that they should freely formulate spontaneous responses as with a human counterpart. By means of exercises and examples, it was explained to them that they could tell as much as they could think of about a question, possibly even beyond the voice assistant's question as the focus was not on the content but on the way the content was reported. Finally, they were played synthesized self-introductions for the two avatar options (s. Figure 4.2) and asked to chose one as an interlocutor during the experiment.

Before the study began, each participant received an introduction into the vehicle controls.

4.1.1.4.2 Phase 2: WoZ Experiment. The drive in the simulator was split into three parts. The first took place with the vehicle parked on a rest area (5 min), followed by the drive on a highway (8 min) and in a city (8 min). Baselines of two minutes without spoken interaction were included at the beginning of the highway and city parts. After the end of the city drive, the lead vehicle parked in a parking bay and thereby indicated the participant to likewise park the vehicle behind it. Both driving sections on the highway and in the city were performed manually with an SAE Level 0 and moderate oncoming traffic. In the city, additional traffic lights were placed approximately each kilometer, which jumped from red to green when the participant approached in order not to influence the traffic flow and to prevent motion sickness.

The spoken interaction consisted of system-initiated questions of the simulated voice assistant to the driver and his or her answers. Depending on the length of a particpant's answers, the voice assistant was able to ask more or less questions.

4.1.1.5 Dependent Variables

Evaluation measures. Different types of data were collected in the course of this user study in order to investigate the linguistic behavior of participants while driving with spoken interaction as a secondary task. Besides the personal information collected in a pre-survey, data was logged in the driving simulator during the participants' interaction with the simulated SDS. These included the WoZ logs and synchronized RTDD logs from which the speech dialog and the respective driving performance could be reconstructed.

Table 4.2 provides an overview of the measures employed in the investigation of language behavior. On the basis of the recorded participant utterances collected in this study, annotations and transcriptions were performed in order to obtain a corpus of spoken language. On this basis, syntax and complexity-related features were computed. In addition, RTDD was recorded during the driving simulation. Here, the driving speed, distance to the vehicle in front, and lane keeping were measured for the highway and city phases. In order to ensure valid interpretations, the logged driving performance measures were reduced to those sequences during which no voice interaction took place (baseline BL) or a user-side response was given on the highway (H) or in the city (C). For the time intervals of these sequences, standard deviations were computed for the driving speed *SPDev* [in km/h], distance to the lead vehicle *DLDev* [in m] and the lateral position on the lane *LPDev* [in m].

Hypotheses. On the basis of these measures, hypotheses were formulated. They are presented in the following and are validated with the statistical analyses results in the following sections.

- In general, voice-based interaction as a secondary task increases a driver's cognitive load at the expense of driving performance. Accordingly, the measured distraction parameters in terms of speed, distance to the lead vehicle and lane keeping were expected to deteriorate during speaking sequences of the drivers compared to driving without voice interaction.

$$SPDev_{\mathrm{BL}} < SPDev_{\mathrm{C+H}} \tag{4.1}$$

Table 4.2: Measures of the user study concerning the examination of language behavior under different driving conditions.

	Measure	Source
Language behavior	Syntactic complexity features (*SC*)	Spoken language corpus from recorded utterances
Driver distraction	Speed deviation (*SPDev*)	RTDD logs
	Deviation of distance to lead vehicle (*DLDev*)	RTDD logs
	Deviation of lateral position (*LPDev*)	RTDD logs

$$DLDev_{\mathrm{BL}} < DLDev_{\mathrm{C+H}} \tag{4.2}$$

$$LPDev_{\mathrm{BL}} < LPDev_{\mathrm{C+H}} \tag{4.3}$$

- This user study was performed in two driving conditions with a different degree of complexity. Compared to steady driving on a highway, urban driving was expected to increase cognitive load. In relation to the different cognitive load with spoken interaction as a constant secondary task, a deterioration of the driving behavior in the city was expected. This included an increased deviation of the driver distraction parameters speed, distance to the driver in front and lane keeping.

$$SPDev_{\mathrm{H}} < SPDev_{\mathrm{C}} \tag{4.4}$$

$$DLDev_{\mathrm{H}} < DLDev_{\mathrm{C}} \tag{4.5}$$

$$LPDev_{\mathrm{H}} < LPDev_{\mathrm{C}} \tag{4.6}$$

- On the basis of the cognitive load induced by driving under different conditions, it is assumed that a driver adapts own language behavior according to his or her cognitive capabilities. It is thus expected that syntactic complexity of a driver's spoken language decreases with an increasing level of driving complexity.

$$SC_{\mathrm{C}} < SC_{\mathrm{H}} \tag{4.7}$$

- As speech reflects human personality, differences in terms of syntactic complexity are expected according to drivers' Big Five personality traits and their assignment to individual user clusters (UC). In this context, human personality is considered as an interplay of the Big Five traits instead of assigning a particular linguistic behavior to one individual

trait.

$$SC_{UC_i} < SC_{UC_j} \tag{4.8}$$

The driving simulation study was conducted according to the described methodology. The following sections describe the construction of a spoken language corpus and extraction of syntactic features as the basis for subsequent investigations concerning linguistic behavior while driving under consideration of different driving conditions and human personality.

4.1.2 Spoken Language Corpus

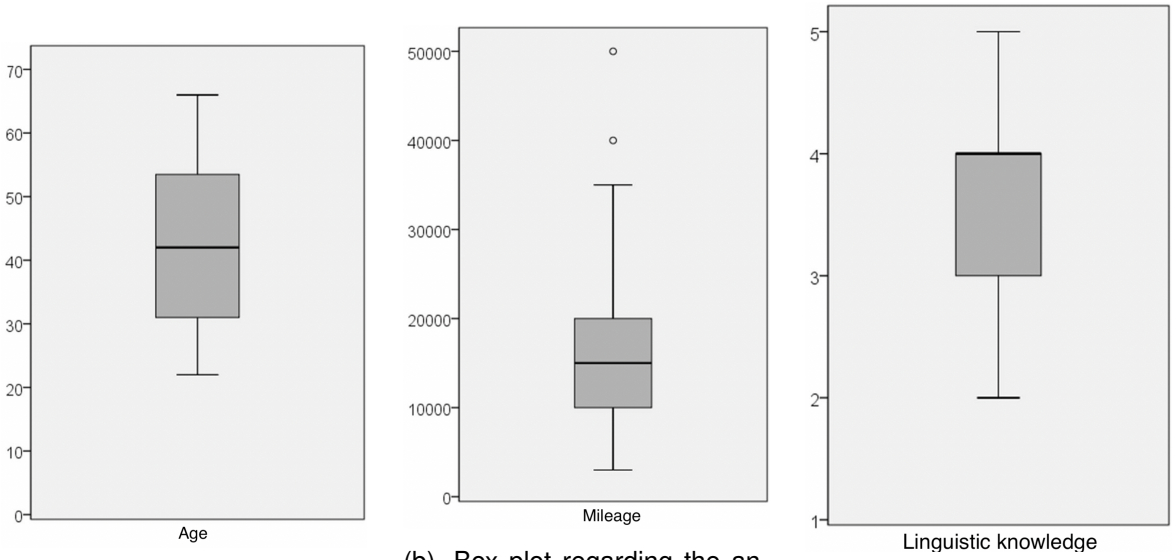
In this section, the collected spoken language data is described. For this purpose, first, the results of the applied pre-survey are summarized. Second, the data pre-processing steps including the transcription and annotation of participants' utterances are described in detail. The resulting corpus served as the basis for analyses of drivers' linguistic behavior with spoken interaction as a secondary task.

4.1.2.1 Questionnaire Results

As introduced in Subsection 4.1.1.3.1, demographic and personal information of the participants was collected by means of a pre-survey.

Overall, the collected data of 47 participants was included in the analyses.[2] Their average age was 42.68 years ($Mdn = 42$, $SD = 13.46$) within a range of 22 and 66 years (s. Figure 4.4a) and a gender distribution of 28 male and 19 female participants. They indicated an average vehicle mileage of 16,893.62 kilometers per year ($Mdn = 15{,}000$, $SD = 10{,}211.32$; s. Figure 4.4b). Further personal information was assessed on 5-point Likert scales. The participants generally rated their linguistic skills as average to good with a mean value of $M = 3.51$ ($Mdn = 4$, $SD = 0.83$, $IQR = 1$, s. Figure 4.4c). Only 8.51% of the participants considered their linguistic knowledge as worse than average. As visualized in Figure 4.5, the self-assessed Big Five personality traits (Rammstedt and Danner, 2016) indicated a homogeneous behavior. In general, the participants considered themselves as tolerable and unselfish

[2] Due to the considerable amount of time required for the manual transcription and annotation, the analyses in this chapter were based on data from only 47 of a total of 72 participants. The speech data of the remaining participants were processed successively and are part of the strategy derivation in Section 5.1.



(a) Box plot regarding age.

(b) Box plot regarding the annual vehicle mileage [in km].

(c) Self-assessed linguistic knowledge.

Figure 4.4: Results of the first user study concerning the age, vehicle mileage and linguistic prior knowledge.

(Agreeableness: $M = 3.78$, $Mdn = 3.80$, $SD = 0.39$, $IQR = 0.50$), as well as disciplined and orderly (Conscientiousness: $M = 4.10$, $Mdn = 4.11$, $SD = 0.47$, $IQR = 0.56$). They were generally open to new experiences (Openness: $M = 3.51$, $Mdn = 3.60$, $SD = 0.43$, $IQR = 0.70$) and extraverted (Extraversion: $M = 3.83$, $Mdn = 3.88$, $SD = 0.56$, $IQR = 0.75$). The participants furthermore considered themselves as low neurotic (Neuroticism: $M = 2.25$, $Mdn = 2.25$, $SD = 0.52$, $IQR = 0.63$). A majority of 34 participants (80.95%) chose Petra (s. Figure 4.2, left) as interlocutor, that is, an avatar with female voice.

4.1.2.2 Transcription and Annotation of Spoken Language Data

Video files were recorded in the driving simulator depicting the road, the user and the vehicle interior. Based on the extracted audio track, the voice recordings were tailored to the individual participant responses ranging from 3 to 400 seconds ($M = 41.78$, $SD = 40.19$) without further processing.

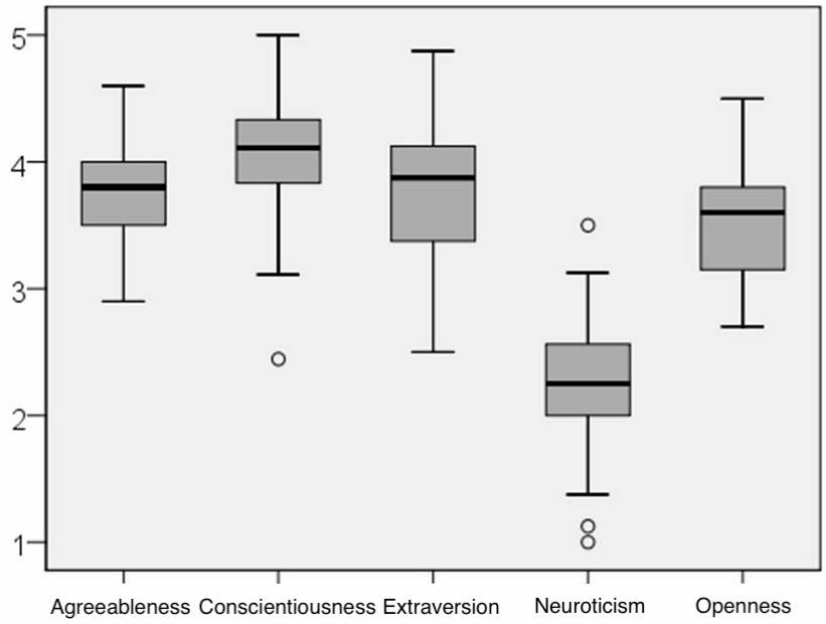


Figure 4.5: Box plot regarding the individual Big Five personality traits according to Rammstedt and Danner (2016) on a 5-point Likert scale.

The spoken participant data was then manually transcribed and annotated in three correction loops with three annotators for the purpose of subsequent Natural Language Processing (NLP). Each annotator thereby followed a specified set of guidelines (s. examples in Table 4.3):

- **Standardization:** For the purpose of automated analysis using publicly available NLP tools, colloquial speech was transferred according to the requirements of written language (Ex. 1).
- **Grammaticality:** In order to avoid failures in the following NLP phase, transcriptions were enhanced to possibly grammatical German sentences. This step required, for example, adding omitted constituents (Ex. 2, 6). Components included by an annotator were excluded from analyses.
- **Redundancy:** In order to reflect user language, comments, repetitions, *etc.*, which do not provide own contributions by the participant were marked. This included, for instance, the (partial) repetition of the original question (Ex. 3) and was excluded from

Table 4.3: Transcription and annotation examples from the German data collection. Adapted from Stier *et al.* (2020c, Table 2) with kind permission from Association for Computing Machinery.

	Transcription	**Annotation**
(1)	*Das ist ne gute Frage.*	Das ist [eine|ne] gute Frage. (eng. "That's a good question.")
(2)	*Und __ ist halt einfach viel schlechter als als schwarzes Leder...*	Und [das|] ist halt einfach viel schlechter als/ als schwarzes Leder... (eng. "And it is simply much worse than black leather...")
(3)	*Mein Lieblingshobby?*	[[Mein Lieblingshobby?]] (eng. "My favourite hobby?")
(4)	*Das ist eine gu-* <pause> *eine gute Frage.*	Das ist (eine gu-//) eine gute Frage. (eng. "That's a good question.")
(5)	*Ich war jetzt* <pause> *gerade* <stutter> *Anfang des Jahres* <pause>*...*	Ich war jetzt// gerade [äh] Anfang des Jahres//... (eng. "I was just now at the beginning of the year...")
(6)	*Ja __ schmeckt lecker wenn sie selber gemacht ist.*	Ja, [Lasagne|] schmeckt lecker, wenn sie selber gemacht ist. (eng. "Yes, lasagna tastes great when it is homemade.")

analyses.

- **Corrections:** A subject's own corrections were marked (Ex. 4). Corrected components were excluded from analyses.
- **Hesitations:** Words of hesitation and pauses were marked as disfluencies (Ex. 5) and were excluded from analyses.
- **Punctuation marks:** Sentence markers were employed according to the requirements of written language to ensure optimal results in the NLP phase (Ex. 6). This includes putting a comma according to the German rules of grammar and a sentence mark to indicate the end of a grammatical sentence.
- **Numbers:** were written out.

Table 4.4: Example answer and a subset of computed features. Taken from Stier *et al.* (2020c, Table 3) with kind permission from Association for Computing Machinery.

Example answer (Subject 44)	**Feature**	**Value**
[[Wenn ich morgen im Lotto [lacht] gewinnen würde?]] [[Ohje ohje]] Also da hab ich jetzt ja schon die ganze Zeit gesagt, ich würde als aller erstes normal das Geld verteilen. Also ([wenn es|wenn's] jetzt) wenn es jetzt [ein|'n]// hoher Lottogewinn ist, dann würd ich das/ jetzt erst mal// meine Eltern schuldenfrei machen, [meinen|mein] Bruder schuldenfrei machen und// [äh] die restliche Familie dann// bedienen. Und// dann würd ich mir irgendwann mal was für mich überlegen. Aber was// hab ich jetzt so// spontan keine/ keine Ahnung.	Proposition count Sentence length Type-token ratio Synt. depth (max) ...	21 68 0.71 5 ...

An annotated example from the spoken language corpus is provided in Table 4.4 (left). Following the procedure described above, the data set consists of 1,037 answers (per participant: $M = 22.06$, $SD = 8.47$), 6,259 annotated sentences ($M = 133.17$, $SD = 39.81$) and 108,056 words ($M = 2{,}299.06$, $SD = 737.24$).

4.1.2.3 Feature Extraction

In a next step, features were extracted from the data set with respect to syntactic complexity. The transcription of spoken language according to the guidelines above was a highly subjective approach, although considered necessary to enable the use of publicly available NLP tools like SpaCy (v2.2, Honnibal and Montani, 2017) and BeNePar (Kitaev and Klein, 2018). Since the subjectively realized concept of a grammatical sentence does not represent a suitable measure for the analyses of user language, average-based features were computed in relation to the number of words in the respective utterance under examination. In total, the following 37 syntax and complexity-related features were computed. An exemplary subset of these computed features is provided in Table 4.4 (right).

- **Dependency labels**[3] (mean) for adverbial components, conjunctions (comparative, conjuncts, coordinating), complementizers, genitive attributes, junctors, modifiers, nega-

[3] `https://spacy.io/api/annotation#dependency-parsing` (last access: 13/02/2020)

tions and phrasal genitives.

- **Dependency length** (max, mean) as the distance from a word to all dependent words.
- **Propositions** (count, mean) as the number of ideas introduced in an utterance (Chand *et al.*, 2012).
- **Root dependency** (max, mean, min) as the number of dependencies a root has (*e.g.*, including auxiliary verbs).
- **Root position** (mean, most left/right) as the positions of the root in an utterance (*e.g.*, shifting due to stopwords).
- **Sentence length** (count) as the number of words in an utterance.
- **Stopwords** (count) as the number of stopwords in an utterance.
- **Syntactic depth** (max, mean) as dependency parse tree depth (Pinter *et al.*, 2016).
- **Syntactic structure** (mean) for the number of complex nominal phrases, prepositional phrases, main clauses, relative clauses, subordinate clauses, subclauses in general and phrase length.
- **Type-token ratio** as the relation of individual word types to the number of words in an utterance.
- **Verb valence** (max, mean) as the number of subject and object realizations of a verb.
- **Word dependency** (max, mean, min) as the number of dependencies a word has.

The corpus described in this section and the linguistic features extracted from it regarding the syntactic complexity of spoken language served as a basis for further analyses of drivers' speech behavior during speech interaction as a secondary task in parallel to driving.

4.2 Syntactic Complexity Components

The set of computed features offers a lot of information regarding the syntactic complexity of the produced language of drivers. However, there is the potential for redundant information, for example, due to high correlations with other features. At the same time, it was necessary to summarize this amount of information to allow for a better interpretation. For this purpose,

an exploratory Principal Component Analysis (PCA) was conducted as the basis for a better interpretability of the features indicating a different linguistic behavior. For this purpose, only the data collected on the highway and in the city were included. Furthermore, the data of five participants had to be excluded since they did not complete both driving conditions due to technical problems in the driving simulator. Therefore, the analysis was performed on a data subset for 42 participants and 665 answers (highway 348, city 317).

Besides the risk of redundant information due to high correlations between the extracted features, a certain degree of correlation between variables is required to perform a PCA. In order to take these requirements into account and to include relevant information only, those features with up to three correlation values $<$.3 or $>$.9 were excluded (Field, 2009). Following this procedure, the features were reduced to a subset of 23 features.

A PCA was conducted on this selected feature set with oblique rotation (SPSS v24.0; promax). The sample size was verified as adequate according to the Kaiser-Meyer-Olkin measure (KMO $=$ .76) and with KMO values for individual items $>$.64. The correlations between items were sufficiently large for a PCA according to Bartlett's test of sphericity ($\chi_2(253) =$ 11,514.83, $p <$ .001). Five components were revealed in an initial analysis with eigenvalues over Kaiser's criterion of 1 and explained 66.13% of variance. Given the sample size and Kaiser's criterion, they were retained for further analysis. Factor loadings after rotation are shown in Table 4.5. Considering the features that cluster on the same sub-components, the following factors were deduced as patterns within the employed data set of spoken language concerning the syntactic complexity of driver language.

Factor 1: General complexity

The features that load highly on this factor all contain some component related to the syntactic complexity of an utterance in relation to its sequential word order. For instance, the syntactic complexity of utterances is assumed to increase with the extension of the distance between dependent lexical units. At the same time, word dependencies and dependency depths are assumed to increase due to the lexical units and their relations enlarging the distance. One prominent feature in the context of this factor is represented by the dependencies a root has. The syntactic complexity of an utterance is assumed to increase with an increasing number of root dependencies and related positions, for instance due to the use of an auxiliary and the thereby adapted positioning of the main verb to verb-last-position.

Factor 2: Lexical complexity

This component summarizes linguistic features which are directly related to the lexical variabil-

Table 4.5: Summary of exploratory factor analysis results indicating rotated factor loadings (N = 665). Adapted from Stier *et al.* (2020e, Table 5) with kind permission from Association for Computing Machinery.

Features	**1**	**2**	**3**	**4**	**5**
Root dependency (mean)	**.937**	-.010	.351	.115	.012
Dependency length (mean)	**-.758**	-.116	.343	-.230	-.013
Root position (mean)	**.749**	.103	-.114	-.152	.358
Word dependency (mean)	**-.717**	-.028	.195	.057	-.023
Syntactic depth (max)	**-.632**	**.500**	-.124	.069	.071
Phrase length (mean)	**.620**	-.006	.248	.053	-.386
Dependency length (max)	**-.536**	.282	.255	-.077	-.023
Junctors (Dep. label)	**.431**	.249	.179	-.163	-.074
Sentence length	.009	**.927**	.023	.063	-.078
Propositions (count)	.041	**.925**	.033	.093	-.074
Type-token ratio	.065	**-.806**	-.135	.055	.062
Root dependency (max)	.223	.291	**.778**	.061	-.015
Word dependency (max)	.015	.276	**.769**	-.012	-.003
Modifiers (Dep. label)	.030	-.119	**.760**	.000	.291
Syntactic depth (mean)	.356	.230	**-.628**	.259	.055
Complementizers (Dep. label)	**-.431**	-.155	.134	**-.672**	.105
Conjuncts (Dep. label)	**-.470**	.029	-.033	**.662**	.079
Coord. conjunctions (Dep. label)	**-.487**	.047	-.054	**.645**	.127
Main clauses (Synt. structure)	.215	-.354	.014	**.507**	.013
Verb valence (mean)	.078	.027	**-.453**	**-.478**	**-.569**
Verb valence (max)	-.027	**.455**	-.166	-.121	**-.518**
Root position (most right)	.164	**.453**	.034	-.279	**.483**
Root position (most left)	**.420**	-.364	.006	-.111	**.423**
Eigenvalues	5.47	4.36	2.27	1.83	1.28
% of variance	23.70	18.94	9.89	7.96	5.56
α	.86	.94	.77	.63	.48

Note: Factor loadings over .40 appear in bold.

ity and information density of an utterance on a superficial level. With an increasing utterance length, the number of lexical units and introduced ideas increases and thereby the utterance's complexity. At the same time, the lexical variability is expected to decrease due to lexical

repetitions of, for example, function words like articles, pronouns or prepositions.

Factor 3: Deep syntactic complexity
This component is related to the parsed dependency depths of an utterance. With an increasing number of modifiers, for instance in the form of additional adjectives or adverbs, the dependencies of the main verb and other part of speech categories, such as nouns, increase. The dependency tree depth is considered here as a proxy for syntactic complexity.

Factor 4: Structural realizations
The features that load on this factor are related to the structural organization of an utterance. Compared to simple, linear main clauses, words, phrases or clauses connected by conjunctions are considered syntactically more complex, for example, by coordinating or subordinating conjunctions. Complementizers here are considered as subordinating conjunctions that introduce clauses with the role of a complement required by the verb.

Factor 5: Realization of verbs and syntactic roles
This component is related to the structural complexity of an utterance in terms of the number of its realized syntactic roles defined by the verb phrase, for example, the subject and possible objects. The syntactic complexity of an utterance is assumed to increase with an increasing number of arguments that a verb requires due to an increased number of relations and information.

Since it was ensured that the features were interrelated, correlations between these extracted components could likewise be observed (s. Table 4.6). In general, the correlation coefficients were rather low. However, trends were observable. As such, notably Factor 2 indicated little relationship with Factor 1 ($r =$ -.03) and Factor 5 ($r =$ .10). Similarly Factors 3 and 4 ($r =$ -.02) and Factors 4 and 5 ($r =$ -.03) were weakly correlated. In contrast, a certain degree of interrelation was observed for Factor 3 with Factor 1 ($r =$ -.18), Factor 2 ($r =$ .19) and Factor 5 ($r =$ .19). Similarly, Factor 2 revealed a relationship with Factor 4 ($r =$ -.20). When interpreting these observations from a linguistic point of view, the complexity of an utterance in terms of the components concerning lexical and verb complexity increased with an increasing deep syntactic complexity. Thus, a relationship was proven between the dependency depth of an utterance and the realization of verb arguments and features like the sentence length and information density. In contrast, as the complexity in terms of the sequential order of individual lexical units increased, a general decrease in dependency depths was observed. At the same time, the structural realization of an utterance in terms of the organization of its clauses and phrases was observed to be related with the lexical complexity of the utterance. With an

Table 4.6: Component correlation matrix. Taken from Stier *et al.* (2020e, Table 6) with kind permission from Association for Computing Machinery.

Factors	**1**	**2**	**3**	**4**	**5**
1	1	-.03	-.18	-.13	.10
2	-.03	1	.19	-.20	-.02
3	-.18	.19	1	-.02	-.19
4	-.13	-.20	-.02	1	-.03
5	.10	-.02	-.19	-.03	1

increasing sentence length and introduced ideas, also the complexity of structural realizations incrased, for instance in terms of a subordinate clause.

On the basis of the syntactic complexity components described above, the following sections present the results of the investigations on syntactic behavior in dependence of the driving situation and a driver's personality traits.

4.3 Syntactic Complexity Under Consideration of Personality and Driving Situation

In this section, the analysis of the linguistic behavior of drivers under consideration of his or her personality traits and the driving situation is described.

4.3.1 Syntactic Complexity and Driving Condition

In order to investigate the linguistic behavior of drivers in relation to the driving condition as described in Stier *et al.* (2020e), an initial analysis was performed to prove the intended effect of differing driving complexities between driving on a highway and in a city with the constant factor of speaking as a secondary task. On this basis, linguistic differences in terms of syntactic complexity between spoken language while driving on a highway or in a city were investigated.

Table 4.7: Driving performance measures and the results of Wilcoxon signed-rank tests (effect size; N = 84). Adapted from Stier *et al.* (2020e, Table 3) with kind permission from Association for Computing Machinery.

	SPDev	***DLDev***	***LPDev***
BL	25.371	25.284	3.566
(H & C)	25.702	22.633	3.834
H	3.472	14.553	0.201
C	2.675	9.701	1.827
BL vs. (H & C)	-1.932 * (.21)	-0.869 (.09)	-3.882 *** (.42)
H vs. C	-2.757 ** (.30)	-3.857 *** (.42)	-5.608 *** (.61)

Note: *$p < .05$, **$p < .01$, ***$p < .001$
Effect size $r = .10$ (small effect), $r = .30$ (medium effect), $r = .50$ (large effect)

4.3.1.1 Driving Complexity and Performance

A first analysis was conducted to demostrate the different degree of driving complexity in this study by means of the assessed RTDD during the combined baseline parts (BL; Baseline I & II in Figure 4.1) and the highway (H) and city (C) phases. As summarized in Table 4.7, driving behavior was directly affected by the parallel secondary task, limited to user-side output in this study due to an increased cognitive load. Wilcoxon signed-rank tests indicated significantly higher values for (H & C) compared to BL in terms of *SPDev* (Z = -1.932, $p = .53$, $r = .21$; BL: 25.371, H & C: 25.702) and *LPDev* (Z = -3.882, $p < .001$, $r = .42$; BL: 3.566, H & C: 3.834). Thus, the increase in cognitive load induced by the task of speaking was reflected in a significant degradation of driving performance in form of an increased deviation from the lane position and driven speed. No significant difference was observed in the case of *DLDev* (BL: 25.284, H & C: 22.633). These findings support Hypotheses 4.1 and 4.3, but are contrary to Hypothesis 4.2.

In a second step, the performance measures for the highway and city phases were compared to demonstrate their different degree of driving complexity. Driving in a city (*SPDev* 2.675, *DLDev* 9.701) showed a significantly lower deviation in speed (Z = -2.757, $p = .006$, $r = .30$) and the distance to the lead vehicle (Z = -3.857, $p < .001$, $r = .42$) compared to driving on a higway (*SPDev* 3.472, *DLDev* 14.553). At the same time, *LPDev* showed a significant increase from the highway to the city (Z = -5.608, $p < .001$, $r = .61$; H: 0.201, C:

1.827). These findings support Hypothesis 4.6, but are contrary to Hypotheses 4.4 and 4.5. A closer look at the participants' driving performance revealed that despite the initial instruction to maintain a distance of 100 m, the distance to the vehicle in front decreased from an average of $M = 84.31$ m on the highway to $M = 50.04$ m in the city. The unexpected results for *DLDev* and *SPDev* are therefore attributed to the fact that by shortening the distance to the vehicle in front, both the general distance and the speed were easier to maintain by a directly visible orientation point. Therefore, in addition to the increased lane deviation in the city, it is concluded from this observation that the cognitive load of the driving task was lower on the highway than in the city. Accordingly, the two driving conditions fulfilled the intended effect of differing driving complexity by inducing a different level of cognitive load. This observation proved the basis for the investigation of language behavior under different driving conditions.

4.3.1.2 Syntactic Complexity Components for Different Driving Conditions

Based on the observation that the two driving conditions in this simulation study exhibit a different degree of complexity, the linguistic behavior of participants while driving was examined with respect to their syntactic complexity by means of the extracted factor components.

The procedure of a PCA allowed the reduction of the original feature set to a smaller subset of interpretable components. The computed factor loads provided the basis for calculating factor scores using the regression method. Here, a Wilcoxon signed-ranks test (s. Table 4.8) comparing factor scores for the highway and the city phases indicated that user language differed in terms of the two components lexical complexity ($Z =$ -2.807, $p = .005$, $r = .11$) and deep syntactic complexity ($Z =$ -2.577, $p = .01$, $r = .10$) due to the differing degree of cognitive load induced by the driving conditions. The provided mean values for the individual factor features revealed that the participants used longer sentences (H: $M = 106.51$, $SD = 105.41$; C: $M = 107.60$, $SD = 103.94$), however, with a lower propositional density (H: $M = 0.339$, $SD = .007$; C: $M = 0.301$, $SD = 0.12$) and lexical variety (H $M = 0.747$, $SD = 0.16$; C: $M = 0.662$, $SD = 0.25$) in the city compared to the highway. Similarly, their utterances were characterized by a lower structural complexity in terms of root dependencies (H: $M = 5.423$, $SD = 1.55$; C: $M = 5.124$, $SD = 2.25$), word dependencies (H: $M = 5.746$, $SD = 1.41$; C: $M = 5.462$, $SD = 2.27$), modifiers (H: $M = 0.242$, $SD = 0.07$; C: $M = 0.224$, $SD = 0.09$) and general syntactic depth (H: $M = 0.266$, $SD = 0.06$; C: $M = 0.237$, $SD = 0.09$).

The results of this analysis have shown that the driving complexity is directly reflected in

Table 4.8: Summary of factors and the results of the comparison between highway and city based on a Wilcoxon signed-ranks test (effect size; N = 710). Adapted from Stier *et al.* (2020e, Table 7) with kind permission from Association for Computing Machinery.

Factor	Features	Highway	City	H vs. C
1		...		-1.324 (.05)
2	Sentence length	106.52 (105.41)	107.60 (103.94)	-2.807 ** (.11)
	Propositions (count)	0.339 (0.07)	0.301 (0.12)	
	Type-token ratio	0.747 (0.16)	0.662 (0.25)	
3	Root dep. (max)	5.423 (1.55)	5.124 (2.25)	-2.577 * (.10)
	Word dep. (max)	5.746 (1.41)	5.462 (2.27)	
	Modifiers (mean)	0.242 (0.07)	0.224 (0.09)	
	Synt. depth (mean)	0.266 (0.06)	0.237 (0.09)	
4		...		-1.052 (.04)
5		...		-0.449 (.02)

Note: * $p < .05$, ** $p < .01$
Effect size $r = .10$ (small effect), $r = .30$ (medium effect)

language use while driving. With an increase in driving complexity, a decrease in the syntactic complexity of a driver's language was observed by means of a number of linguistic features. This finding generally confirms Hypothesis 4.7.

4.3.2 Syntactic Complexity and User Personality

In this section, based on Stier *et al.* (2020c), the analysis of the linguistic behavior of drivers under consideration of their personality is described. For this purpose, user clusters were constructed based on the participants' self-assessed Big Five personality traits. On this basis, linguistic differences in terms of syntactic complexity in the spoken language between these user clusters were investigated.

Table 4.9: Summary of Big Five trait cluster centroids (SD) and additional descriptive information. Adapted from Stier *et al.* (2020c, Table 6) with kind permission from Association for Computing Machinery.

		UC 1	**UC 2**	**UC 3**	**UC 4**	**UC 5**	**UC 6**
Subjects		8	6	6	9	5	13
Age		41.88	37.83	41.33	44.11	43.00	44.92
Gender	male (%)	10.64	4.26	10.64	10.64	8.51	14.89
	female (%)	6.38	8.51	2.13	8.51	2.13	12.77
Ling. competence		3.50	3.50	3.00	3.33	4.00	3.69
Veh. mileage/year (K)		12.88	13.83	19.17	19.78	24.60	14.77
Big Five trait	**Openness**	3.625 (0.46)	3.750 (0.21)	2.900 (0.18)	3.122 (0.12)	3.920 (0.38)	3.707 (0.23)
	Conscientious.	3.917 (0.46)	3.741 (0.79)	3.852 (0.22)	4.481 (0.27)	4.644 (0.32)	3.929 (0.30)
	Extraversion	3.703 (0.70)	3.354 (0.31)	3.542 (0.38)	3.806 (0.37)	4.725 (0.14)	3.759 (0.39)
	Agreeableness	4.150 (0.22)	3.167 (0.19)	3.533 (0.15)	4.067 (0.41)	3.540 (0.30)	3.829 (0.12)
	Neuroticism	2.750 (0.23)	2.854 (0.37)	2.458 (0.29)	1.917 (0.48)	1.800 (0.62)	2.036 (0.26)

4.3.2.1 Big Five Personality Clusters

Human personality manifests multiple traits simultaneously. Thus, instead of investigating traits individually at this point, a two-step cluster analysis was conducted using the entire set of self-assessed Big Five personality traits (s. Subsection 4.1.1.3.1) of all 47 participants as clustering variables. The results of the obtained six cluster solution (SPSS v24.0, average silhouette 0.4, cluster size ratio 2.15) are summarized in Table 4.9.

Overall, each user cluster (UC) can thus be characterized by a varying constellation of Big Five traits. UC 1, for example, was characterized by comparatively high levels of Neuroticism ($M = 2.750$, $SD = 0.23$) and Agreeableness ($M = 4.150$, $SD = 0.22$). UC 2 differed from this by a lower manifestation of the trait Agreeableness ($M = 3.167$, $SD = 0.19$) and a compar-

atively higher Openness to new experiences ($M = 3.750$, $SD = 0.21$). Although participants assigned to UC 3 also exhibited comparable levels of Neuroticism ($M = 2.458$, $SD = 0.29$), they differed from the former user clusters in being comparatively less open ($M = 2.900$, $SD = 0.18$), conscientious ($M = 3.852$, $SD = 0.22$), and extraverted ($M = 3.542$, $SD = 0.38$). In contrast, these traits were manifested in UC 6 as differentiating attributes (Openness: $M = 3.707$, $SD = 0.23$; Conscientiousness: $M = 3.929$, $SD = 0.30$; Extraversion: $M = 3.759$, $SD = 0.39$). Overall, UC 4 and UC 5 showed the highest values in terms of Conscientiousness (UC 4: $M = 4.481$, $SD = 0.27$; UC 5: $M = 4.644$, $SD = 0.32$) and Extraversion (UC 4: $M = 3.\ 806$, $SD = 0.37$; UC 5: 4.725, $SD = 0.14$) of the participants and, at the same time, the lowest neurotic expression (UC 4: $M = 1.917$, $SD = 0.48$; UC 5: $M = 1.800$, $SD = 0.62$). In particular, they could be distinguished by the degree of Openness to new experiences (UC 4: $M = 3.122$, $SD = 0.12$; UC 5: $M = 3.920$, $SD = 0.38$).

4.3.2.2 Syntactic Complexity Components for Personality Clusters

On the basis of the identified user clusters according to the participants' Big Five personality traits, the linguistic behavior of participants while driving was examined in order to investigate differences in terms of syntactic complexity by means of the extracted factor components.

The PCA procedure allowed the assignment of the originally extracted features to more interpretable components. As indicated in Subsection 4.3.1.2, factor scores were calculated for the individual data points using a regression method based on the factor loadings. In this way, differences between the syntactic complexity in spoken language was examined under consideration of the participants' personality. Here, Kruskal-Wallis tests for independent samples (s. Table 4.10) comparing factor scores according to the individual user clusters indicated that user language differed in terms of the components general complexity (Factor 1, $H(5)=$ 43.056, $p < .001$), lexical complexity (Factor 2, $H(5)=$ 18.823, $p = .002$), and the realization of verbs (Factor 5, $H(5)=$ 15.723, $p = .008$). Subsequent post hoc tests using Bonferroni correction (s. Table 4.11) revealed that concerning the general complexity component, particularly the user clusters UC 1, UC 4 and UC 6 differed from the user clusters UC 2, UC 3 and UC 5.[4] Taking the respective feature values in Table 4.10 into account, the spoken language of UC 1, UC 4 and UC 6 was characterized by a higher syntactic complexity compared to the re-

[4] For reasons of readability, the individual test statistics and feature values are not included at this point and reference is made to their presentation in Tables 4.10 and 4.11.

Table 4.10: Results of the comparison between user clusters based on a Kruskal-Wallis test ($H(df = 5)$; N $= 707$) and overview of feature values (SD).

Factor	Features	UC 1	UC 2	UC 3	UC 4	UC 5	UC 6	Test
1	Root dep. (mean)	0.230 (0.056)	0.228 (0.082)	0.243 (0.076)	0.227 (0.072)	0.238 (0.079)	0.217 (0.068)	43.056***
	Dep. length (mean)	3.181 (0.406)	3.175 (0.669)	3.048 (0.466)	3.152 (0.494)	3.077 (0.502)	3.147 (0.438)	
	Root pos. (mean)	0.015 (0.008)	0.018 (0.013)	0.019 (0.010)	0.016 (0.009)	0.017 (0.010)	0.015 (0.008)	
	Word dep. (mean)	0.819 (0.034)	0.802 (0.059)	0.805 (0.045)	0.809 (0.036)	0.807 (0.047)	0.816 (0.035)	
	Synt. depth (max)	7.450 (2.238)	6.860 (3.231)	6.850 (2.417)	7.280(2.367)	6.690 2.159)	7.720 (2.543)	
	Phrase length (mean)	0.757 (0.025)	0.753 (0.047)	0.749 (0.033)	0.760 (0.030)	0.752 (0.034)	0.748 (0.026)	
	Dep. length (max)	19.840 (7.827)	18.550 (11.176)	17.370 (9.753)	19.590 (9.446)	15.600 (6.238)	19.280 (9.143)	
2	Sentence length	128.200 (93.978)	91.840 (104.202)	100.040 (120.893)	136.120 (121.871)	82.790 (45.791)	122.930 (105.252)	18.823**
	Propositions (count)	36.040 (26.365)	25.360 (29.196)	27.720 (29.282)	36.950 (35.194)	22.200 (12.494)	34.440 (28.599)	
	Type-token ratio	0.729 (0.101)	0.791 (0.129)	0.773 (0.117)	0.735 (0.135)	0.784 (0.102)	0.739 (0.101)	
3	...	...						11.065
4	...	...						4.672
5	Verb valence (mean)	0.185 (0.043)	0.173 (0.067)	0.186 (0.047)	0.194 (0.048)	0.172 (0.053)	0.176 (0.049)	15.723**
	Verb valence (max)	2.660 (0.756)	2.370 (0.676)	2.440 (0.697)	2.620 (0.639)	2.470 (0.638)	2.520 (0.628)	

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

maining user clusters indicating an averagely higher number of word dependencies, a greater syntactic depth and dependency lengths. At the same time, they revealed a lower average of root positions when compared with UC 2, UC 3 and UC 5. When comparing the cluster centroides (s. Table 4.9), the main difference between these two cluster groups was found in the manifestation degree of the Big Five trait Agreeableness (comparably low: UC 2, UC 3, UC 5; comparably high: UC 1, UC 4, UC 6). In terms of their lexical complexity, particularly UC 4 and UC 6 differed significantly. Here, the spoken language of UC 4 indicated generally longer sentences and a higher information density measured as introduced propositions, while the lexical variability by means of the type-token ratio was slightly lower compared to UC 6. The cluster centroids of these two user clusters particularly revealed a difference in terms of the manifested degree of the Big Five trait Openness (comparably low: UC 4; comparably high: UC 6). Additionally, for UC 6 a significantly deviating realization of verbs was observed compared to UC 3. Here, UC 6 showed a higher maximum verb valency but in general a lower average verb valency in relation to the words per utterance than UC 3. The main difference between these user clusters was again observed for their degree of Openness (comparably low: UC 3; comparably high: UC 6).

Overall, particularly UC 1 and UC 4 revealed higher feature values and were thus identified with a comparably high syntactic complexity. At the other end of the continuum were UC 2, UC 3, and especially UC 5, which had comparatively lower feature values. Their syntactic complexity was therefore estimated to be comparatively lower.

The results of this analysis have shown that the personality of a driver is directly reflected in spoken language use as a secondary task while driving. Accordingly, the identified user clusters differ in terms of linguistic features of syntactic complexity and thus can generally be differentiated. However, the transitions between user clusters are rather fluid and language behavior cannot be as clearly assigned to them as was the binary case for driving complexity (s. Section 4.3.1). Nonetheless, these findings generally confirm Hypothesis 4.8.

4.3.3 Discussion of Results

This section summarizes and discusses the results of the analyses concerning the linguistic behavior of drivers under consideration of the driving situation and the Big Five personality traits. A summary of the respective hypotheses and their validation is provided in Table 4.12.

Table 4.11: Results of pairwise Kruskal-Wallis tests indicating linguistic differences between users clusters UC 1 to UC 6 ($H(df = 5)$ (effect size); N $= 707$). Based on Stier *et al.* (2020c, Table 7) with kind permission from Association for Computing Machinery.

Clusters	**(1)** General complexity	**(2)** Lexical complexity	**(5)** Realization of verbs
UC 1 vs. UC 2	4.905*** (.13)	0.737 (.02)	-2.199 (.06)
UC 1 vs. UC 3	3.363* (.09)	0.597 (.02)	-2.909 (.08)
UC 1 vs. UC 4	0.461 (.01)	-1.917 (.05)	-0.357 (.01)
UC 1 vs. UC 5	3.83** (.10)	0.801 (.02)	-1.659 (.04)
UC 1 vs. UC 6	0.979 (.03)	2.164 (.06)	-0.175 (.00)
UC 2 vs. UC 3	-1.769 (.05)	-0.183 (.00)	-0.474 (.01)
UC 2 vs. UC 4	-4.456*** (.12)	-2.46 (.07)	1.862 (.05)
UC 2 vs. UC 5	-1.102 (.03)	0.04 (.00)	0.548 (.01)
UC 2 vs. UC 6	-4.404*** (.12)	1.12 (.03)	2.208 (.06)
UC 3 vs. UC 4	-2.886 (.08)	-2.464 (.07)	2.537 (.07)
UC 3 vs. UC 5	0.632 (.02)	0.231 (.01)	1.074 (.03)
UC 3 vs. UC 6	-2.712 (.07)	1.452 (.04)	2.999* (.08)
UC 4 vs. UC 5	3.376* (.09)	2.569 (.07)	-1.316 (.03)
UC 4 vs. UC 6	0.465 (.01)	4.236*** (.11)	0.217 (.01)
UC 5 vs. UC 6	-3.241* (.09)	1.108 (.03)	1.63 (.04)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$, based on Bonferroni correction
Effect size $r = .10$ (small effect), $r = .30$ (medium effect)

- The results of an initial analysis show that the task of speaking as a secondary task increases the cognitive load of a driver compared to driving without voice-based interaction. For this purpose, the three driving performance measures *SPDev*, *DLDev* and *LPDev* were compared between baseline phases and the sequences, where the participants of this user study answered the questions of a simulated voice assistant. In this context, *SPDev* and *LPDev* were significantly higher when the participants were speaking compared to the combined baselines. Thus, a degradation of the participants' driving performance was observed in the form of an increased deviation from the lane position and driven speed. These results coincide with the general consent that driving

Table 4.12: Validation of Hypotheses.

$SPDev_{BL} < SPDev_{C+H}$	✓	(4.1)
$DLDev_{BL} < DLDev_{C+H}$	(✗)	(4.2)
$LPDev_{BL} < LPDev_{C+H}$	✓	(4.3)
$SPDev_{H} < SPDev_{C}$	(✗)	(4.4)
$DLDev_{H} < DLDev_{C}$	(✗)	(4.5)
$LPDev_{H} < LPDev_{C}$	✓	(4.6)
$SC_{C} < SC_{H}$	✓	(4.7)
$SC_{UC_i} < SC_{UC_j}$	✓	(4.8)

performance is influenced by the secondary task of language production and support Hypotheses 4.1 and 4.3. In the case of *DLDev*, no statistically significant result was revealed. For this reason, Hypothesis 4.2 has to be rejected.

- The analyses described in this section furthermore indicate that the driving conditions on a highway and in a city included in this user study differ in terms of their driving complexity and the induced cognitive load of the participants while driving and speaking. A significantly increased deviation in the lateral position on the lane *LPDev* during the city indicated a degradation in driving performance compared to the highway. This finding supports Hypothesis 4.6. Although *SPDev* and *DLDev*, contrary to the expectation, showed significantly lower values when driving in the city compared to the highway, the inspection of these parameters indicated a clear deviation from the initial instruction to maintain a distance of 100 m to the lead vehicle. While the average distance to the lead was 84.31 m on the highway, it decreased to 50.04 m in the city. This shortening of distance can be attributed to the increased cognitive load of the city, which made both distance and speed easier to maintain and may thus explain the unexpected results for *SPDev* and *DLDev*. However, since no clear conclusion about the driving behavior of the participants is obtained through the results at this point, the Hypotheses 4.4 and 4.5 are rejected.
- On the basis of the above observations, the results of a linguistic analysis demonstrate that the degree of driving complexity is directly reflected in the produced language of a driver. More precisely, differences in the linguistic behavior of participants were disclosed while driving on a highway or in a city due to the different cognitive demand of

the driving task. Based on a PCA for the purpose of better interpretability of syntactic complexity features and the resulting five complexity components, which characterize the spoken language of the participants, the factor scores derived from the factor loads via regression were compared. Here, a difference in the linguistic behavior between the highway and the city was particularly identified for features that highly load on components related to the lexical complexity and the deep syntactic complexity of an utterance. Overall, a lower syntactic complexity was found in the produced language of participants while driving in the city in comparison to their linguistic behavior on the highway. Thus, a general simplification of the participants' language for the city was observed in comparison to the highway. This finding confirms Hypothesis 4.7. It can therefore be concluded that drivers appear to adapt their syntactic behavior in these terms to the respective driving situation and according to their cognitive abilities.

- The results furthermore indicate that the affiliation to a user cluster based on the interplay of the Big Five traits is reflected in the produced language of a driver. Based on a comparison of the factor scores computed from the five complexity components within a PCA, it was observed that the complexity in terms of the general syntactic complexity component in spoken language increased with an increasing degree of agreeableness (the "agreeable clusters": UC 1, UC 4, UC 6; the "less agreeable clusters": UC 2, UC 3, UC 5). Moreover, for the trait Openness as a differentiating factor between the user clusters, it became apparent that with a lower manifestation (the "less open clusters": UC 3, UC 4; the "open cluster": UC 6) the syntactic complexity increased in terms of the lexical complexity component and the realization of verbs. These findings generally confirm Hypothesis 4.8. A reason why these three syntactic complexity components including relatively shallow features provide clearer results in the differentiation of user clusters than the PCA components with features of a deeper analysis (as in the deep syntactic complexity and structural realization components) may be due to the reliability of the NLP tools in use and in their application to the very specific data set of drivers' small talk answers. Nevertheless, the results coincide with the general consent that the personality of a user is directly reflected in language style and that therefore personality traits can be differentiated based on linguistic features. The results further confirm that this differentiation based on syntactic complexity features is also applicable to user utterances in the interaction context of a secondary task.

4.4 Summary of Results

This section provides a summary of the user study and analyses on language production. In a first step, the data collection study in a driving simulator and the main results are summarized. Finally, implications on the following research are outlined.

4.4.1 Summary

In this section, a driving simulation study was presented against the background of collecting linguistic data of drivers while driving at an SAE level 0 and constructing a spoken language corpus (s. Section 4.1). For this purpose, the participants interacted with a simulated SDS in two driving conditions and answered small talk questions by the simulated voice assistant. The collected spoken data was then manually transcribed and annotated in three annotation loops with three annotators on the basis of a specified set of guidelines. In a next step, syntactic and complexity related features were extracted using publicly available NLP tools. On this basis, an exploratory PCA was conducted for the purpose of better interpretability of the features (s. Section 4.2). Following this procedure, five syntactic complexity components were revealed, which characterize the participants' linguistic behavior from a general and lexical complexity over the deep syntactic complexity to structural realizations and verb valency. On their basis, the linguistic behavior of participants was compared under consideration of the driving situation and the driver's Big Five personality traits (s. Section 4.3). For this purpose, in a first step speaking as a secondary task was proven to be more complex by inducing a higher cognitive load on the driver measured by his or her driving performance compared to driving without voice-based interaction. In a second step, the task of driving on a highway and in a city as two different driving situations was investigated and a generally higher cognitive load in the city was identified compared to the highway. Thus, driving in a city with traffic lights was found to be generally more demanding and thus more complex than driving on a highway. The subsequent analysis of linguistic behavior under consideration of the driving condition (s. Section 4.3.1) demonstrated that this degree of driving complexity was directly reflected in the produced language of a driver by means of a general simplification in terms of its lexical and deep syntactic complexity in the city compared to the highway. It was thus proven that a driver adapts own syntactic behavior to the respective driving situation and according to own cognitive abilities. Therefore, the influence of the interaction context in the form of a driving sit-

uation plays a crucial role in voice-based interaction and should be taken into account. For the purpose of investigating the produced language of participants while driving with respect to their personality traits (s. Section 4.3.2), a two-step cluster analysis based on the participants' self-assessed Big Five personality traits revealed six user clusters with different manifestation degrees of the individual traits. A comparison of the five syntactic complexity components between them revealed that the interplay of the Big Five traits was directly reflected in the syntactic behavior of drivers in terms of the produced general complexity, lexical complexity and verb valency. Here, particularly the manifestation of the traits Agreeableness and Openness were revealed as differentiating factors between the six user clusters. In general, the results are thus in line with the common agreement that the personality of an SDS user is directly expressed in linguistic style. The results of this section furthermore demonstrated that the differentiation of human personality traits based on syntactic cues can also be applied in the interaction context of a secondary task.

Overall, both the driving situation and a driver's personality were identified as factors influencing the use of syntactic features while driving as a primary task. Thereby, the research work presented in this section provides a further step towards the development of an adaptive strategy in dual-task environments with respect to the syntactic design of voice output. The findings regarding the form of spoken language and its influencing factors can be combined in a next step with the previous conclusions concerning syntactic preferences in the perception of in-vehicle voice output (s. Section 3).

4.4.2 Implications on Research Work

The results of this sections have shown that a driver's personality and the driving situation are reflected in language production as a secondary task in terms of syntactic features. Therefore, regardless of the strategy (*cf. attraction* vs. *complementarity* principle) in a dual-task environment, such as driving a car, syntactic complexity features need to be taken into account when designing in-vehicle SDS voice output. At this point, in addition to the user-side speech production, the influence of language perception as secondary task evidently needs to be considered (s. Section 3). A next step described in the following chapter therefore involves investigating the applicability of adaptation principles. As an example, the question will be examined whether a rather extraverted user prefers an "extraverted" SDS that reflects his or her language style – or an "introverted" SDS with a complementary language behavior. Here,

one prerequisite for a user- and situation-adaptive SDS is the automated differentiation and identification of human personality traits and the respective driving condition by analyzing the output of a driver. However, the approach of speaker and situation recognition does not match the focus of this work. For this reason, a pilot experiment is briefly described and discussed in the following chapter, while more in-depth research is referenced to future work.

Chapter 5

Development of an Adaptive Dialog Strategy

This chapter describes the development of a user- and situation-adaptive strategy concerning the syntactic form of voice output in SDS interaction as a secondary task in the dual-task scenario of driving. In this context, the need to consider individual users and situations does not match the conventional development cycles of an SDS, where a user experience expert extensively defines and evaluates a system's functionality (*e.g.*, including possible input, semantic scope, dialog strategy, system reaction or voice output) for a 'stereotyped' user, who may rarely exist in reality (Hjalmarsson, 2005b). Instead, the findings and observations of the previous chapters (s. Chapters 3 and 4) are used as a baseline in order to propose a development approach and to subsequently deduce a voice output strategy: Speech-based interaction in the vehicle as a secondary task generally imposes the requirement that speech needs to be produced and processed in parallel with driving. The influence on and of the primary task therefore represents a central factor with regard to successful, intuitive communication. In accordance with prior research, it was observed in this work that the cognitive load of a driver measured in terms of driving performance increases when listening to voice output or when the driver produces own utterances. For the development of a voice output strategy, it is therefore considered essential to take the driver-SDS-interaction in the context of driving into account. Furthermore, in terms of speech perception, the exploratory investigation of the perceived naturalness and comprehensibility of in-vehicle voice output revealed a direct influence of the syntactic form of prompts on both the cognitive driver load and subjective preferences and its relation to various user and system parameters. Here, a general preference for simple syntactic forms as opposed to nested, more complex phrases was observed.

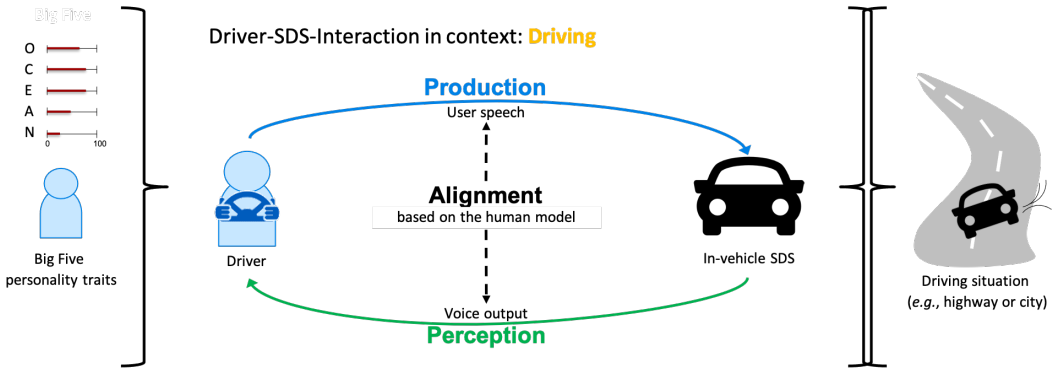


Figure 5.1: An adaptation strategy for voice output is developed in the interaction context of driving under consideration of the produced driver language and the perception of in-vehicle voice output by the driver.[1]

Given that a person's personality manifests itself directly in his or her speech behavior, the Big Five personality traits were used next to the driving situation as a framework for distinguishing users and characterizing a driver's spoken language while driving on the speech production side. Overall, particular syntactic complexity components were identified, which serve as indicators for the manifestation of particular personality traits and the current driving complexity. Following the concept of linguistic alignment according to the model of interpersonal communication and taking into account that a driver was observed to adapt the syntactic complexity of own utterances according to his or her cognitive capabilities with respect to the current driving situation, the adoption of the *similarity principle* concerning voice output generally seems to be a reasonable choice: If in a complex driving situation the complexity of a driver's speech decreases, this linguistic behaviour should be reflected in system outputs in order to keep the cognitive load as low as possible and to avoid increasing it unnecessarily. The extent to which this assumption holds for the dual-task vehicle context and which role the driver's own personality and the preferred system behavior play in this context is addressed in this chapter.

Thus, the focus here is on the use of language in user utterances and system-side voice output from a user perspective. By combining the two aspects of speech perception and speech production as visualized in Figure 5.1, by relating a driver's perception of voice output and indicated preferences to the linguistic properties of spoken driver language, both individual requirements of the SDS user and the driving context are to be included.

In this chapter a development approach is proposed and a user- and situation-adaptive strategy concerning the syntactic design of voice output is presented in Section 5.1. On this theoretical basis, the realization of the developed adaptation strategy is described in Section 5.2 by means of a prototype implementation. Finally, Section 5.3 presents the evaluation of the adaptation strategy by comparing the prototipical realization with a non-adaptive WoZ within the framework of a real-life user study in the car. The results of this chapter are then summarized and discussed in Section 5.4.

5.1 Development Approach and Strategy Deduction

In this section, an approach for the development of a user- and situation-adaptive strategy for the syntactic form of in-vehicle SDS voice output is presented. For this purpose, the findings of this research work concerning the aspects of spoken language perception in terms of in-vehicle voice output (A; s. Chapter 3) and a driver's speech production (B; s. Chapter 4) are combined in order to ensure a holistic approach from a user perspective in the context of HMI as a secondary task following the human model (C; s. Figure 5.1). The here proposed approach is based on the findings and data collected in the driving simulation studies presented in the previous chapters and builds on the publication by Stier *et al.* (2020a). In general, it comprises the following steps, which are further explained in the subsequent sections:

A Speech Perception: Gather driver preferences for in-vehicle voice output in the form of syntactic paraphrases, that is, with a comparable semantic complexity but a different syntactic realization, based on its perceived naturalness and comprehensibility (details in Section 5.1.1).

B Speech Production: Characterize driver speech by means of syntactic complexity features on the basis of a collection of spoken language while driving as an extract of actual language behavior (details in Section 5.1.2).

C User- and Situation-Adaptive Strategy: Derive a strategy by comparing the specified preferences (A) with actual language behavior (B) under consideration of the personality of a driver in the specific interaction context of driving (s. Section 5.1.3).

[1] Car icon made by Freepik from `https://www.flaticon.com/`; online: 24/03/2021.

5.1.1 Speech Perception (A)

In this context, the collected user preferences for two syntactic variants (MCV, RCV) of 46 participants (27 male, 19 female) with an average age of 41.98 years ($SD = 15.07$) as described in Section 3.3 are reemployed. In summary, the syntactic paraphrases were assessed 368 and 1,104 times with regard to their perceived naturalness and comprehensibility, respectively, within three explanation types (What, How, When), two domains (COP, DAS) and two driving complexities (AUT, MAN).

5.1.2 Speech Production (B)

The spoken language corpus described in Subsection 4.1.2 including the transcriptions of all study participants is used for this purpose. It thus comprises the spoken data of 72 German native speakers (46 male, 26 female) with an average age of 42 years ($SD = 13.28$). Only considering the data collected during the simulation drives, that is, on the highway or in the city, the corpus contains 1,220 answers (per participant: $M = 16.94$, $SD = 5.83$), 5,706 annotated sentences ($M = 79.25$, $SD = 38.46$) and 97,543 words ($M = 1{,}354.76$, $SD = 719.05$). As described in Subsection 4.1.2.3, a total of 37 syntax and complexity related features were extracted on this basis. For the purpose of their better interpretability and in order to reduce redundant information, first features with less than four correlation values $>.3$ and $<.9$ were excluded (Field, 2009). Thereby, the features were reduced to a subset of 19 features (s. Table 5.1) and an exploratory PCA (SPSS v24.0) with oblique rotation (oblimin) was conducted in order to deduce syntactic complexity patterns, which characterize user language while driving. The sample size was verified as adequate (KMO$=.81$), with all KMO values for individual items $>.66$. The correlations between items were sufficiently large (Bartlett's $\chi_2(190) = 25{,}179.220$, $p < .001$). Four components were revealed with eigenvalues over 1 according to Kaiser's criterion. Factor loadings after rotation are shown in Table 5.1. Further details of the PCA are provided in Appendix C.1. Considering the features that cluster on the same sub-components, the following factors were deduced as patterns within the spoken language data:

Factor 1: Surface complexity

The features that highly load on this factor all contain some component related to the syntactic complexity of an utterance on a superficial level, including aspects such as lexical variability

Table 5.1: The results of an exploratory factor analysis indicating rotated factor loadings (N = 1,220). Based on Stier *et al.* (2020a, Table 1),

Features	**(1)**	**(2)**	**(3)**	**(4)**
Sentence length	**.958**	.004	-.018	-.032
Stopwords (count)	**.955**	.003	-.020	-.032
Propositions (count)	**.951**	.035	-.013	-.046
Type-token ratio	**-.799**	-.002	-.186	-.007
Syntactic depth (max)	**.617**	**-.507**	-.195	.019
Root position (most right)	**.596**	**.564**	.085	.118
Verb valence (max)	**.484**	-.058	.155	.011
Dependency length (max)	**.430**	-.350	.316	.100
Root position (mean)	.033	**.891**	-.131	-.011
Word dependency (mean)	-.035	**-.755**	.062	-.030
Root dependency (mean)	-.173	**.722**	.280	-.283
Dependency length (mean)	-.007	**-.606**	.385	.283
Root position (most left)	-.346	**.425**	-.124	.043
Root dependency (max)	.288	.110	**.753**	-.169
Word dependency (max)	.283	-.075	**.743**	-.025
Modifiers (Dep. label)	-.125	.035	**.731**	.010
Syntactic depth (mean)	.040	.217	**-.721**	-.211
Complementizers (Dep. label)	-.068	-.034	.049	**.846**
Relative clauses (Synt. structure)	-.260	-.102	-.138	**.828**
Main clauses (Synt. structure)	-.365	-.194	-.137	**-.749**
Eigenvalues	6.86	3.57	1.89	1.63
% of variance	34.30	17.87	9.43	8.15
α	.90	.82	.81	.76

Note: Factor loadings over .40 appear in bold.

and length correlations. For instance, with an increasing utterance length, the number of lexical units increases and likewise the occurrence of stopwords and introduced ideas. At the same time, the lexical variability is expected to decrease due to repetitions of, for example, function words. The interaction of these factors is related to the increased syntactic complexity of an utterance, which equally appears to be reflected in its maximum dependency tree depth, verb valence, and distance of dependent lexical units.

Factor 2: General complexity
The features of this component are related to the syntactic complexity of an utterance in relation to its sequential word order and dependencies between individual units. With regard to the roots of an utterance, the syntactic complexity is assumed to increase with an increasing number of root dependencies and related positions, for example, by means of auxiliary verb constructions including a main verb shift to verb-last-position. At the same time, the syntactic complexity is expected to increase with the number of dependants a word has and the related extension of the distance between dependent lexical units.

Factor 3: Deep structural complexity
This factor is related to the dependency structures of an utterance. One prominent feature in the context of this component is represented by the number of modifiers, such as adjectives or adverbs, which in turn appears to be related with the maximum number of lexical dependencies. Since modifiers are directly linked to their reference word, the use of such modifiers does not automatically lead to an increased dependency depth as, for example, in the case of inserted prepositional phrases. Regardless of how it is derived, the dependency tree depth is considered a proxy for the syntactic complexity of an utterance, that is, the deeper, the more complex.

Factor 4: Structural realizations
The features that load on this factor are related to the structural organizations of an utterance. In this context, syntactic complexity increases with the use of complementizers as subordinating conjunctions and nested relative clauses. At the same time, the number of simple, linear main clauses is expected to decrease.

There is no single, universal feature to describe syntactic complexity; it reveals itself in different ways. For this purpose, a large number of features was computed to reflect syntactic complexity and assigned to interpretable components in the context of this PCA. The relationship of these features can also be observed in the correlation among the resultant factor

Table 5.2: Component correlation matrix.

Factors	**1**	**2**	**3**	**4**
1	1	-.12	.34	.06
2	-.12	1	-.11	-.23
3	.34	-.11	1	.13
4	.06	-.23	.13	1

components (s. Table 5.2). Although the correlation coefficients were generally rather low with values between 0.6 and 0.34, trends can be observed. Thus, in the case of Factor 1, a low association with Factor 4 ($r = .06$) was found, whereas the correlation degrees with Factor 2 ($r =$ -.12) and Factor 3 ($r = .34$) were considerably higher. Similar coefficients were indicated for the relationship between Factors 2 and 3 ($r =$ -.11), Factors 2 and 4 ($r =$ -.23) and Factors 3 and 4 ($r = .13$). From a linguistic perspective, this indicated a comparably low correlation between surface complexity measures and the structural realizations of an utterance. Thus, surface features such as sentence length, type-token ratio, and number of propositions (Factor 1) were only marginally related directly to the syntactic form of a sentence (Factor 4). For instance, the length of a sentence conveyed only limited information about its syntactic realization. In contrast, there was a stronger correlation between these surface measures and dependency-related components at the level of individual lexical units (Factor 2) and phrases or sentences (Factor 3), since individual lexical items form the basis for higher-level phrases, which in turn can fulfill syntactic roles at the sentence level and thus are interrelated. For instance, with an increasing sentence length and number of propositions, likewise the general complexity on the basis of lexical units appeared to decrease, while with an increasig number of lexical material in an utterance the need for deeper dependency structures increased, for instance by the occurence of additional phrasal modifiers and an overall increased syntactic dependency depth. Similarly, the general complexity on the basis of lexical units (Factor 2) was only conditionally interrelated with the deeper structural complexity on phrase and sentence level (Factor 3). The more complex a sentence was structured, the simpler the lexical dependency relations appeared, for example, by shifting modifiers from the lexical to the sentence level through the use of subordinate clause constructions and the accompanying distribution of dependencies across several components. Likewise, this explains the positive relationship of structural complexity (Factor 3) with structural realizations (Factor 4).

The above observations are generally consistent with the complexity factors and included features derived in Chapter 4. The major difference is found in the combination of Factors 1 and 5 regarding the realization of verbs and verb valence (s. Section 4.2) into Factor 1 (s. Table 5.1). Since the above analysis is based on a significantly expanded data set as a reliable resource (N= 1,220) compared to the initial analysis and identification of potential complexity factors in Chapter 4 (N= 665), it will be taken as a reference in the following sections.

The components identified in this section regarding the syntactic complexity of spoken language in the context of interaction as a secondary task during driving serve as the basis for deriving an adaptation strategy for voice output in the following.

5.1.3 User- and Situation-Adaptive Strategy (C)

The approach presented here aims at combining the aspects of speech perception and production within voice-based interaction as a secondary task. For this purpose, the user preferences concerning syntactic forms (s. Section 5.1.1) are directly put in relation with the syntactic characteristics in driver language (s. Section 5.1.2). In order to additionally take the individual user in the respective interaction context of driving into account, the approach of user clusters according to study participants' self-assessed Big Five traits is reemployed. For this purpose and in order to obtain a more comprehensive understanding of drivers' personality traits in this context, the self-assessments of the previous user studies (s. Subsections 3.2.1.1 and 4.1.1.1) were combined. As described in Subsection 4.3.2.1 (p. 144), human personality in this research work is interpreted as an interplay of the Big Five traits. Therefore, instead of separating and investigating the traits individually, they were treated as a whole and all of them served simultaneously as clustering variables within a two-step cluster analysis (SPSS v24.0). Following this procedure, three user clusters were obtained from the data of 118 included participants (average silhouette 0.4, cluster size ratio 2.15). In a next step, the syntactic preferences (s. Section 5.1.1) were extracted for each user cluster and driving complexity. Accordingly, complexity scores were computed for each factor component from the rotated factor loads f (s. Table 5.1) and the extracted standardized syntactic complexity feature values v (s. Subsection 4.1.2.3) according to Formula 5.1.

$$\text{complexity score} = \frac{\sum_{i=1}^{n-1} f_i * v_i}{\sum_{i=1}^{n-1} f_i} \tag{5.1}$$

On this basis, the linguistic behavior of each user cluster and driving situation converted into complexity scores was classified according to a binary scale as either “complex” or “simple.” For this purpose, the respective complexity scores were compared with the corresponding cross-cluster averages of the respective driving complexity under the premise that a higher value was considered as more simple and a lower value as more complex. The background for this evaluation is justified by the relation of the complexity values calculated for the voice prompts employed in the user study in Section 3.3 (s. Appendix A.4.3) as a basis for comparison (s. Appendix C.2).

Table 5.3 summarizes the results of the described procedure and further provides additional descriptive information for the individual user clusters.[2] All share a similar gender distribution of approximately 60% male and 40% female. With a mean age of 35.76 years ($SD = 10.06$), the members of UC 2 are younger than UC 1 ($M = 41.38$, $SD = 15.22$) or UC 3 ($M = 45.22$, $SD = 12.62$) members. Overall, UC 1 is distinguished from the other user clusters by the comparatively highest scores regarding the Big Five traits Openness ($M = 3.37$, $SD = 0.42$), Conscientiousness ($M = 4.\ 23$, $SD = 0.38$), Extraversion ($M = 4.09$, $SD = 0.42$), and Agreeableness ($M = 3.93$, $SD = 0.35$), and the comparatively lowest score for Neuroticism ($M = 2.05$, $SD = 0.44$). Whereas UC 2 differs from UC 3 in the manifestation degree of its members with regard to the components Openness ($M = 3.69$, $SD = 0.40$), Extraversion ($M = 3.76$, $SD = 0.52$) and Neuroticism ($M = 3.38$, $SD = 0.47$), members of UC 3 can be characterized as comparatively more conscientious ($M = 3.98$, $SD = 0.53$) and agreeable ($M = 3.79$, $SD = 0.49$). According tho their most outstanding characteristics, the members of UC 1 are summarized as the “extroverts & conscientious”, whereas UC 2 can be referenced as the “neurotics & open” and UC 3 as the “conscientious& agreeable.”

The last row in Table 5.3 indicates the deduced adaptation strategy by inspecting whether the classified complexity scores prevailingly reflect or contrast the participants’ syntactic preferences: Mirroring the results concerning syntactic preferences from Section 3.3.2, all user clusters indicated a clear preference for the syntactically simpler MCV ($M = 4.39$ on a 5-point Likert scale) compared to the more complex RCV ($M = 4.17$) in either driving situation complexity of a highway (MCV: $M = 4.44$, RCV: $M = 4.29$) or city (MCV: $M = 4.41$; RCV: $M =$

[2]The user clusters are based on the combined 118 participants of the studies on speech perception (46 participants; s. Section 5.1.1) and speech production (72 participants; s. Section 5.1.2). The syntactic preferences are extracted from the study on speech perception, that is, including the combined 1,472 assessments concerning the perceived comprehensibility and naturalness of 46 participants. The computation of complexity scores is based on 1,220 spoken language transcriptions of 72 participants taken from the study on speech production.

Table 5.3: Derivation of an adaptation strategy for in-vehicle voice output under consideration of the Big Five trait user cluster centroids and the driving situation complexity. Based on Stier *et al.* (2020a, Table 2),

User cluster		**UC 1**		**UC 2**		**UC 3**		**All**	
Number of subjects		58		33		27		118	
Mean age (SD)		41.38 (15.22)		35.76 (10.06)		45.22 (12.62)		40.79 (13.98)	
Gender	male (in %)	60.34		66.67		59.26		61.86	
	female (in %)	39.66		33.33		40.74		38.14	
Big Five trait (SD)	Openness	3.73 (0.42)		3.69 (0.40)		3.03 (0.57)		3.48 (0.50)	
	Conscientious.	4.23 (0.38)		3.57 (0.53)		3.98 (0.53)		3.93 (0.54)	
	Extraversion	4.09 (0.42)		3.76 (0.52)		3.15 (0.46)		3.67 (0.59)	
	Agreeableness	3.93 (0.35)		2.58 (0.40)		3.79 (0.49)		3.43 (0.78)	
	Neuroticism	2.05 (0.44)		3.38 (0.47)		2.57 (0.47)		2.67 (0.74)	
Driving complexity		**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**
Preference	MCV (s)	4.64	4.63	4.04	4.00	4.63	4.58	4.44	4.41
	RCV (c)	4.54	4.46	3.88	3.54	4.44	4.46	4.29	4.15
Complexity scores	Factor (1)	0.6588 (c)	0.7452 (s)	0.6027 (c)	0.5722 (c)	0.8500 (s)	0.8177 (s)	0.7038	0.7117
	Factor (2)	0.8123 (s)	0.6796 (c)	0.8720 (s)	0.7898 (s)	0.6918 (c)	0.7597 (s)	0.7920	0.7430
	Factor (3)	0.7468 (c)	0.8085 (c)	0.8544 (s)	0.8485 (s)	0.7068 (c)	0.8681 (s)	0.7693	0.8417
	Factor (4)	0.5908 (c)	0.6396 (c)	0.9146 (s)	1.0406 (s)	0.5843 (c)	0.6937 (c)	0.6966	0.7913
Adaptive statregy		**contrast**	**contrast**	**mirror**	**mirror**	**contrast**	**mirror**	(c)	(s)

Note: H – Highway with low driving complexity; C – City with high driving complexity; MCV – Main clause variant; RCV – Relative clause variant; (s) – “simple” syntactic behaviour; (c) – “complex” syntactic behaviour

Information on sample sizes: Syntactic preferences based on 1,472 assessments (UC 1: 896; UC 2: 96; UC 3: 480). Complexity scores based on 1,220 spoken language samples (UC 1: H 257, C 212; UC 2: H 291, C 286; UC 3: H 94, C 80).

4.15). In contrast, the binary classified complexity scores vary per user cluster from, for example, predominantly complex in the case of UC 1 ($3x$ complex, $1x$ simple) to predominantly simple in the case of UC 2 ($1x$ complex, $3x$ simple). In this way, by relating the syntactic preference to these syntactic complexity classifications, the derived strategy for user- and situation-dependent voice output involves the application of different adaptation principles for each user cluster and driving complexity. In this respect, a general complementary syntactic behavior is suggested in the case of UC 1, since the reported user preferences of this cluster in favor of simple syntactic forms contrast with the syntactic language behavior classified as predominantly complex in both driving situations. Similarly, a mirroring behavior is suggested for UC 2. In the case of UC 3, the driving situation and inherent complexity represents the decisive factor between the application of either a complementary syntactic behavior for simple, or a mirroring behavior for complex driving situations. Broken down to the user clusters identified above, the adaptation strategy indicates that “extraverted” users (UC 1) prefer a contrasting, and therefore introverted, SDS with a complementary language style. Similarly, “introverted” users (UC 2) seem to prefer an SDS with similar personality traits that reflects their linguistic behavior. Consequently, “agreeable” users (UC 3) prefer a different language style depending on the driving complexity in order to accommodate their respective needs.

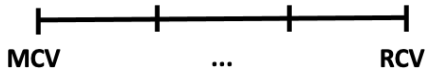
Overall, the user preferences broken down by user clusters show that voice output in the driving context tends to be uniformly preferred as syntactically simple so that it is perceived as natural and comprehensible as possible. As straightforward as this finding seems, in addition to the characterization of spoken language with different driver personalities and driving situations, on a theoretical level it led to the conclusion that the use of one adaptation principle (*e.g.*, *similarity* or *complementarity* principle) in the context of driver-SDS-interaction as a secondary task is not applicable as a universally valid strategy. Rather, the choice of one or the other principle for the goal of intuitive and naturally perceived voice output depends on both individual user and situation characteristics.

An interpretation of the derived strategy at cluster level and its meaning for the form of voice prompts is provided in Table 5.4. It illustrates which speech behavior of a driver is expected in the respective driving situation and which voice prompt’s complexity level follows according to the adaptation strategy. While the derivation of the general adaptation strategy is based on the classification of complexity values in relation to the cross-cluster average of the respective driving situation, the attribute of syntactic complexity clearly has its own meaning per individual user cluster. For instance, the spoken language of UC 1 members in an urban driving environ-

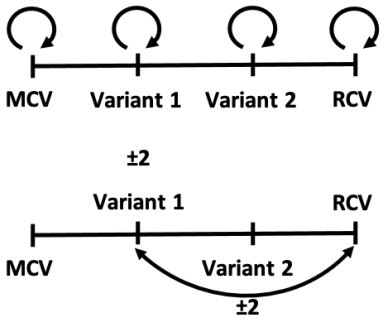
Table 5.4: Interpretation of the deduced adaptation strategy.

Driving context	**Cognitive load**	**Driver utterance**			**Voice output**		
		UC 1	**UC 2**	**UC 3**	**UC 1**	**UC 2**	**UC 3**
highway	low	complex	simple	complex	simple	simple	simple
city	high	simple	complex	simple	complex	complex	simple

ment is generally estimated as complex ($3x$ complex, $1x$ simple) in an inter-cluster comparison of complexity scores with UC 3 ($1x$ complex, $3x$ simple); hovewer, it is considered as simple compared to the spoken language on a highway in a within-cluster comparison under the premise that, as before, a higher value is interpreted as "simple" and a lower value as "complex" (H: $3x$ complex, $1x$ simple; C: $1x$ complex, $3x$ simple). In this context, Table 5.4 indicates that the original assumption of drivers adapting their spoken language while driving needs to be revised. While for UC 1 and UC 3 the assumption applies that a driver in a cognitively demanding driving situation like a city adapts his or her syntactic language behavior according to the induced cognitive load and compensates with comparatively simple syntactic forms, the speech of UC 2 members is estimated as more complex in the city compared to the highway with a generally lower cognitive load. Overall, according to the adaptation strategy described above, the respective syntactic behavior is mirrored or contrasted in SDS voice output leading to generally simple syntactic forms with one exception for UC 1 and UC 2 in the driving context of a city. On this basis, the general adaptation strategy was refined against the background of optimizing user experience and driving safety: As described above, drivers adapt their language to the driving situation according to their cluster affiliation. For example, UC 1 members tend to use more complex syntactic forms in situations with low driving complexity, such as the highway, since less mental resources are occupied by the primary task of driving and are thus free for unrestricted language behavior. In contrast, in a more cognitively demanding situation such as the city, where more mental resources are reserved for the primary task, they tend to use a comparatively simpler syntactic complexity. Consequently, members of this user cluster compensate for the increased cognitive load of the city by using simpler language. It is therefore assumed that, at least for UC 1, a complex, additionally cognitively demanding voice output as envisaged by the adaptation strategy in an already cognitively demanding driving situation is not user-friendly and detrimental to driving safety. To solve this problem and to



(a) The sentence types MCV and RCV are considered as endpoints of a syntactic complexity continuum with possible intermediate levels.



(b) Mirror (above) and contrast (below) strategies for the syntactic complexity continuum comprising two additional sentence types as intermediate levels between MCV and RCV.

Figure 5.2: Refinement of the adaptation strategy.

bring the concept of adaptive voice output closer to human language behavior, the until then binary construct of syntactic complexity (simple vs. complex) was extended. Here, the previously considered sentence types MCV and RCV are understood as poles of a continuum of syntactic complexity with possible intermediate levels (s. Figure 5.2a). For this purpose, in this research work two additional sentence types are vicariously considered.[3] Although this approach only partially represents the in principle infinite possibility of syntactic structures in human language, it is used to approximate SDS voice output to realistic human speech behavior that does not only occur at the level of syntactic complexity of main or relative clauses, but can take complexity levels between these endpoints by combining variable coordinating and subordinating operations. Following this approach, Figure 5.2b depicts the refined strategy for adapting voice output according to the complexity of user speech. While the mirror principle provides that a user utterance is followed by a voice output of similar syntactic complexity, the contrast principle involves matching an utterance's complexity with a prompt of ± 2 complexity levels provided a four-level scale. In this way, the application of the contrast principle prevents a syntactically simple user utterance from being answered with complex, cognitively demanding voice output, as in the case of UC 1 in the city, and unnecessarily increasing the cognitive load of the UC 1 driver.

[3] Details can be found in Section 5.2.

The here described adaptation strategy was implemented and evaluated in the context of a real-life user study. The following sections describe the details of the realization and prototype implementation, and subsequent evaluation.

5.2 Realization

This section describes the realization of the deduced and refined adaptation strategy presented above. First, Section 5.2.1 describes the prototype implementation within the framework of a JavaScript (JS) websocket (WS) application. Second, Section 5.2.2 provides details concerning the input of external contextual information to the prototype. A detailed description of the evaluation of the adaptation strategy by means of the prototypical realization is presented later in Section 5.3.

5.2.1 Prototype Implementation

Against the background of this work, the focus of the prototype was on analyzing the syntactic complexity of a user utterance, such as a response to a small talk question, and providing a syntactically adapted explanation regarding a vehicle function (hereinafter referred to as domain QAS) according to the developed adaptation strategy. An overview of the prototype realization is presented in Figure 5.3, which runs as an offboard JS WS application. The framework was provided by Cerence Studio[4] as a sample application. In Figure 5.3 the grey parts correspond to the utilized Cerence components, while the green parts represent module extensions for the realization of the aforepresented adaptation strategy. External input is colored in blue.

Automatic Speech Recognition (ASR): As described in Section 2.1, Cerence's ASR component captures a user's utterance and translates the speech signal into text.

Natural Language Understanding (NLU): The NLU component analyzes the most probable ASR hypothesis and produces a semantic interpretation of the user utterance. For the purpose

[4] Cerence Studio is a web-based development environment for the design and realization of speech-based applications. For instance, it allows for the definition of dialog flows and the training of own NLU models. See `https://cerence.com/cerence-products/cerence-studio` (Online: 24/04/2021)

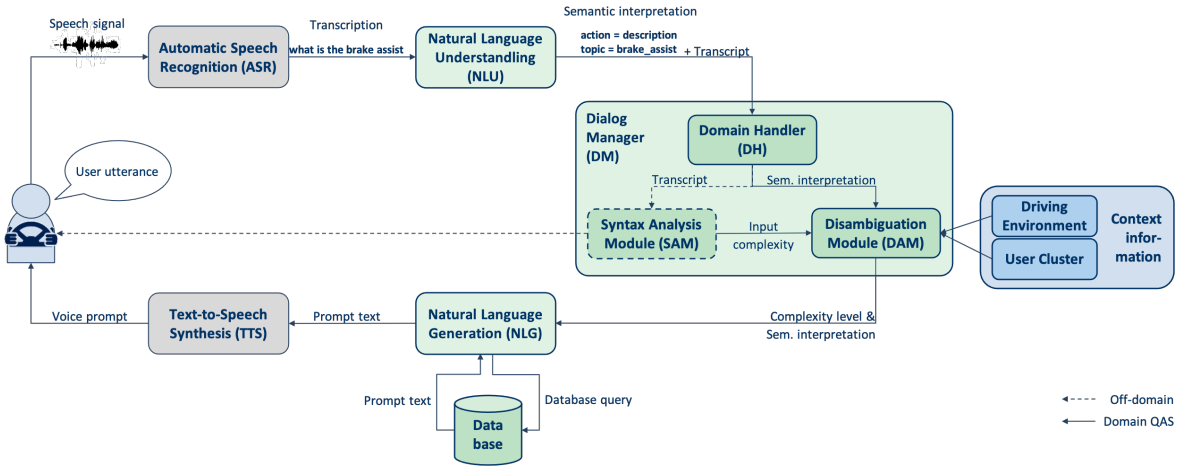


Figure 5.3: Prototype realization within an extended SDS architecture (module extensions in green, external input in blue).

of this work, an NLU model was built and trained employing the Cerence Studio[5] capability to capture user requests within the domain QAS demanding for a vehicle function's explanation.

Dialog Manager (DM): The DM of this prototype consists of an upstream Domain Handler (DH), and a subsequent Disambiguation Module (DAM) and Syntax Analysis Module (SAM). In a first step, the DH identifies a user utterance based on its semantic interpretation as within or out of the domain QAS. For an off-domain utterance, such as in the case of a user answering a small talk question, the SAM analyzes the ASR transcript and stores the resulting syntactic complexity level. The dialog ends at this point and the prototype is waiting for further user input. For an utterance identified as within the QAS domain, the DAM requests the recently identified input complexity level from the SAM and translates it into a targeted output complexity level according to the adaptation strategy under consideration of the user cluster and driving situation. It hands both the concept values of an utterance's semantic interpretation and the complexity level to the next component.

Natural Language Generation (NLG): The NLG component defines the system response that is prompted to the SDS user (s. Section 2.1). In the prototype of this work, the task of the NLG module consists of a database query by means of the DM's output, that is a semantic interpretation combined with a syntactic complexity score. By matching these parameters, a

[5]See footnote 4.

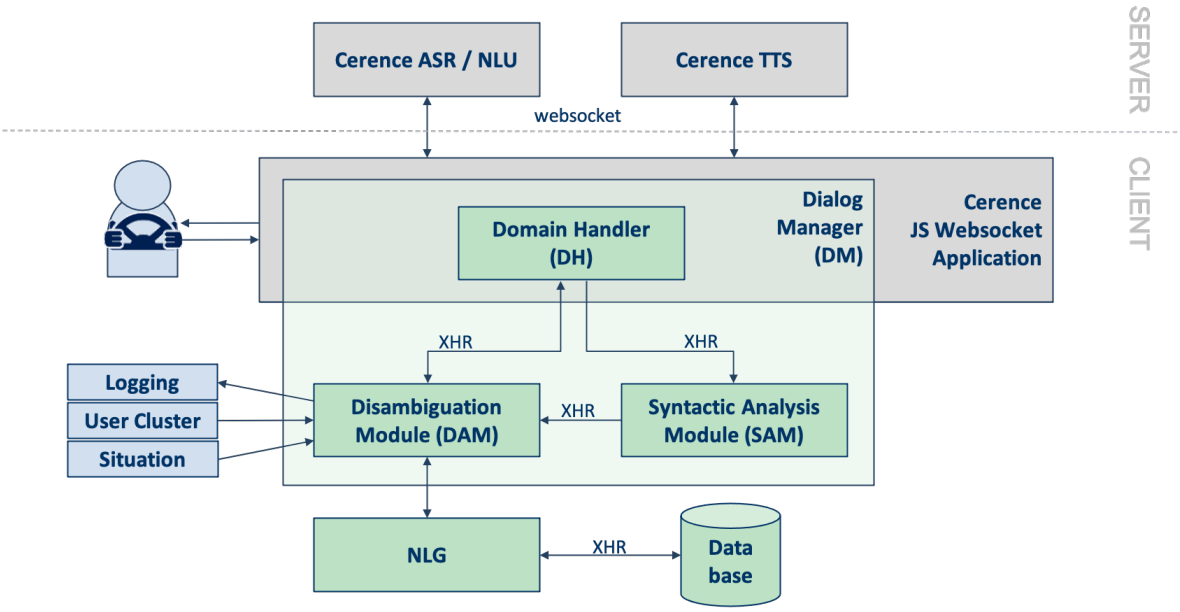


Figure 5.4: Extended prototype architecture as part of a JavaScript websocket application (module extensions in green, external input in blue).

prompt text is retrieved from a database of syntactic paraphrases.

Text-to-Speech Synthesis (TTS): The TTS component transforms the selected answer text into a speech signal (s. Section 2.1). In the prototype of this thesis, system responses were synthesized using Cerence's Petra-ML voice.

Figure 5.4 illustrates details of the prototype architecture as part of a JS WS application. Here, the WS application with the DM assumes the role of a client and communicates via a WS protocol with the ASR, NLU and TTS components on WS servers. It listens for incoming and emerging events, such as a user input and system response, and accordingly triggers the respective components. As introduced before, the DM consists of the DH, DAM and SAM as three individual modules. While the DH acts on the basis of incoming or outgoing event messages and is architecturally assigned to the WS application as an intermediary, communication outside the WS application with and between the DAM and SAM takes place sequentially via XMLHttpRequests (XHR). In order to define the form of a voice prompt, based on which the NLG component queries a voice output, the DAM receives additional input from external sources regarding the affiliation of the SDS user to a user cluster and the respective

situation context. Details of this are described in Section 5.2.2. In order to evaluate the adaptation strategy developed in this work in a user study, the individual steps of the DAM are additionally logged.

The following sections provide details concerning the individual module extensions of the prototype realization.

5.2.1.1 Natural Language Understanding

For the context of the prototype in this work, a custom NLU model was built and trained using the facility provided in Cerence Studio. For simplicity, a single NLU concept `function_type` was modeled for this purpose referring to the vehicle function queried in a QAS. In order to be able to account for any `function_type` value in the form of synonymous uses of a vehicle function, a Daimler-internal survey was conducted by means of a dedicated survey tool developed in-house. It was based on six vehicle functions, which are presented in detail in the subsequent user study in Section 5.3. As in Section 3.3, the participants were given a vehicle function in combination with a question type with the request to formulate a free query based on this information. A total of 14 employees participated in the anonymous survey. Table 5.5 summarizes the included vehicle functions and the additional synonyms considered by the NLU.

5.2.1.2 Domain Handler

Based on the semantic interpretation of the NLU, the DH decides whether a user utterance falls within or outside the domain QAS. In case the NLU concept `function_type` is mapped with a valid attribute value as defined in Table 5.5, a user input is considered as within the domain and the DAM is triggered. In contrast, for example, when the SDS user answers a small talk question, the ASR string cannot be mapped to an NLU concept in the semantic interpretation phase, so the SAM is activated in the case of off-domain utterances.

5.2.1.3 Syntactic Analysis Module

In this prototype, SAM is triggered by the DH in case a user utterance is identified as out of the domain QAS, such as in the case of an answer to a small talk question. The task of SAM is

Table 5.5: The NLU model includes a total of six vehicle functions and synonymous uses based on an anonymous Daimler-internal data collection.

Vehicle function	**function_type**	**Question type**	**Synonyms**
Aufmerksamkeits-Assistent (Attention Assist)	`attent_assist`	what	Attention Assist, Attention Assistent, Aufmerksamkeits-Assi, Aufmerksamkeits-Ding, Einschlafwarner, Müdigkeits-warner
Beduftungssystem (Perfume Atomizer)	`perf_atom`	what	Beduftung, Parfüm
Brems-Assistent (Brake Assist)	`brake_assist`	how	Brems-Assi, Brems-Assistent, Bremskraftverstärker, Dreieck im Außenspiegel
Totwinkel-Assistent (Blind Spot Assist)	`blind_assist`	how	Totwinkel-Assi, Tote-Winkel-Assistent
Sprach-Assistent (Voice Assist)	`voice_assist`	how	Linguatronic, SDS, Sprachansagen, Sprach-Assi, Sprachbedienung, Spracherkennung, Sprachsteuerung, Sprachsystem
Massage-Funktionen (Massage functions)	`mass_funct`	which	Massage-Arten, Massage-Formen, Massage Functions, Massage-Möglichkeiten, Massage-Optionen, Massage-Programme, Massage-Systeme, Massage-Typen, Massagen

then to identify the syntactic complexity of the user utterance based on the ASR output string and to define a syntactic complexity level. The underlying idea here is to compute different complexity measures for the user utterance and align them with those defined within the strategy development in Table 5.3 (p. 164). For this purpose, SAM follows a similar procedure as described in Section 5.1.3 to compute complexity scores: In a first step, syntactic complexity

features are computed with the help of the python-based NLP library SpaCy (v2.2, Honnibal and Montani, 2017). The 20 features summarized in Table 5.1 (p. 159) are employed as the base of operations in this context, since they have been identified as valuable measures to characterize the syntactic complexity of spoken language while driving. Subsequently, the computed features are standardized by means of the respective feature mean and standard deviation values of the collected corpus described in Section 4.1.2.3 as a representative sample of in-vehicle spoken language. This step ensures comparability between the complexity scores of the user utterance and the scores defined within the strategy development, which serve as the basis for comparison. In a second step, four factor scores are computed according to the factor components identified in Section 5.1.2 using Formula 5.1 (p. 162) and the factor loads presented in Table 5.1 (p. 159). Finally, the complexity scores of the user utterance are employed to derive a syntactic complexity level of the user input (s. Table 5.6). The procedure chosen for this purpose follows a naive approach by comparing the user's complexity scores with the identified characteristics of syntactic behavior on the highway and in the city as described in Table 5.3 (p. 164). By means of this simple framework, the closest match per factor needs to be identified in order to classify a driver's spoken language with respect to its syntactic complexity. This classification lays the foundation to, in a later step in the DAM, determine voice output with a corresponding complexity in relation to the adaptation strategy.

It should be noted in this context that the term 'complexity level' is not to be interpreted as judgmental in terms of a uniform cross-cluster complexity since the assignment to one level by means of abstract complexity scores does not allow for a direct mapping to a more or less syntactic complexity across user clusters. Thus, level 1 does not automatically imply a simpler syntactic complexity than level 2 in a direct comparison of user clusters: For instance, as described in Table 5.3, members of UC 2 generally utter syntactically more complex sentences on the highway than in the city, whereas members of UC 1 use simpler syntactic structures on the highway. In this regards, the complexity levels as defined here serve as a simple categorization per user cluster onto a four-level scale of syntactic variants including MCV and RCV as its extremes.

5.2.1.4 Disambiguation Module

The DAM is triggered by the DH in case a user utterance is identified as within the domain QAS, that is when the user is asking a question related to some in-vehicle functionality. The

Table 5.6: The complexity scores computed as syntactic characteristics of spoken language on the highway (H) and in the city (C) are mapped to a syntactic complexity level and serve as a framework to classify the syntactic complexity of a user utterance.

User cluster	**Complexity scores**	**Complexity level**			
		1	**2**	**3**	**4**
UC 1	Factor (1)	0.6588	0.6804	0.7020	0.7452
	Factor (2)	0.8123	0.7459	0.7128	0.6796
	Factor (3)	0.7468	0.7623	0.7777	0.8085
	Factor (4)	0.5908	0.6030	0.6152	0.6396
UC 2	Factor (1)	0.6027	0.5875	0.5798	0.5722
	Factor (2)	0.8720	0.8309	0.8104	0.7898
	Factor (3)	0.8544	0.8514	0.8500	0.8485
	Factor (4)	0.9146	0.9461	0.9776	1.0406
UC 3	Factor (1)	0.8500	0.8338	0.8258	0.8177
	Factor (2)	0.6918	0.7088	0.7257	0.7597
	Factor (3)	0.7068	0.7471	0.7875	0.8681
	Factor (4)	0.5843	0.6116	0.6390	0.6937
		$\widehat{=}$ H	–	–	$\widehat{=}$ C

task of DAM is then to retrieve the required information in order to decide on the syntactic complexity level of a system response, which is assembled by the NLG component. Three sources of information form the central role in this decision, which are the syntactic complexity of user speech, the affiliation of a user to a personality cluster and the current driving situation. In the scenario of a conversational dialog system, continuous spoken HMI is assumed. The complexity level of prior user utterances is therefore considered. Whereas the latter is requested from the SAM, both the user cluster and driving situation are gathered from external sources. Although a detailed investigation on how this context information can automatically be derived from in-vehicle HMI is out of the scope of this work, Section 5.2.2 will present the results of a pilot experiment in order to demonstrate and hint at possible future approaches. In this prototype, context information is retrieved by means of configurable settings, which are manipulable at runtime. Under consideration of these factors, DAM deduces the syntactic complexity of the system answer to be prompted following the adaptation strategy defined in Table 5.3 (p. 164) and its refinement in Figure 5.2 (p. 167) and forwards it to the NLG component. By default,

in case SAM cannot provide any complexity level, *e.g.*, due to ASR failure or because a user did not utter any suitable speech, the sentence type MCV is selected as not to run the risk of overloading the driver with a syntactically complex system response in the worst case.

The DAM additionally provides a logging of the ASR transcript, its semantic interpretation, the syntactic complexity level indicated by the SAM and its decision for a syntactic complexity of the output prompt.

5.2.1.5 Natural Language Generation

The NLG component selects an output prompt for a QAS according to the complexity level and semantic interpretation provided by the DAM. This task is accomplished in form of a database query using `PostgreSQL`[6], where a predefined SQL query is employed by the prototype as canned string and filled with the respective parameters to be accessed.

5.2.1.6 Pilot Study: The Reliability of ASR Transcripts

The reliability of the prototype to compute a syntactic complexity score based on ASR strings and the selected linguistic features at runtime was investigated in a small experiment. For this purpose, 25 subjects with an average age of 34.43 years ($SD = 11.08$) and a gender distribution of 18 male and 7 female were invited to interact with the prototype in a fixed-base driving simulator at the Daimler site in Sindelfingen following the experimental setup of Section 4.1 and a study design, which will be described in detail in Section 5.3: In summary, the participants answered small talk questions by a simulated voice assistant as linguistic input material for the prototype to compute a syntactic complexity score. Subsequently, the participants were triggered via visualized tasks on the head-unit to formulate one-shot, in-vehicle related questions, which were answered by the simulated voice assistant with explanatory voice prompts in a particular syntactic form as defined by the prototipical realization of the adaptation strategy. Overall, each participant thereby experienced between 8 and 10 QAS, as dependent on the length of the participants' small talk answers a varying number of QAS could be performed within the fixed experiment duration of 16 minutes.

Following the experiment, the recorded small talk answers of the participants were transcribed and annotated following the procedure defined in Section 4.1.2.2. On this basis,

[6] `https://www.postgresql.org` (last access: 27/12/2021)

syntactic complexity features were extracted as described in Section 5.1.2 and fed into the prototypical realization of the adaptation strategy to automatically retrieve syntactic complexity levels from SAM and accordingly a syntactic output complexity by DAM. The reliability of raw ASR strings as input for the prototype was then estimated by comparing the complexity levels computed by the prototype at runtime in the experiment and the complexity scores computed offline based on the manually transcribed and annotated voice recordings of the participant's speech.

The results of this experiment showed that the prototype failed to compute any reliable ASR candidate string in 26.86% of interactions due to either connectivity issues or because the employed ASR engine was not entirely stable enough to resolve spontaneous driver speech. Although this observation represents a further limitation of the prototype, the ASR performance as such represents a given dependency, which cannot be improved within the scope of this work. Thus, more importantly, the results furthermore demonstrated that the complexity scores based on the raw and annotated ASR strings coincided in 68.93% and only deviated in 31.07%. These results led to the conclusion that, in the context of this research, the reliability of the prototype to compute the syntactic complexity of spontaneous spoken language based on raw ASR transcripts is considered as reliable enough for further investigations with a focus on the adaptation strategy itself.

5.2.2 Driving Situation and User Cluster

As visualized in Figure 5.3, both the driving situation and affiliation of a user to a particular personality cluster are characterized as external factors influencing the selection and prompting of voice output. From an architectural point of view, this type of contextual information is collected based on the respective user and situation, and processed by the DAM in order to define the syntactic complexity of the prototype's voice output. This dependency is visualized in Figure 5.5.

The relevance of user personality and driving situation has been sufficiently illustrated in the course of the present work. However, an automated identification of the affiliation to a user cluster or the respective driving situation based on HMI in a dual-task scenario are considered beyond the scope of this thesis. For this reason, the prototypical realization of user- and situation-adaptive voice output is based on the input on client-level (s. Figure 5.4). The following section describes the approach of the prototype (s. Subsection 5.2.2.1). Nonethe-

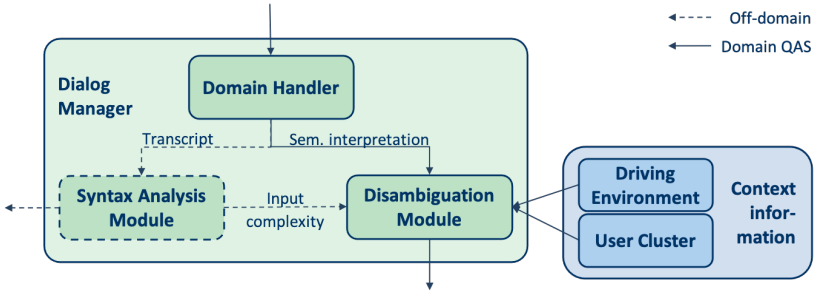


Figure 5.5: Context information, such as the user cluster and driving situation, is processed by the Disambiguation Module (DAM) in order to define the syntactic complexity of the prototype's voice output according to the developed adaptation strategy.

less, a pilot study has been performed in order to prove the feasibility to automatically detect personality and driving situation from user speech in the context of driving as primary task. The results of this experimental approach are presented in Subsection 5.2.2.2.

5.2.2.1 Contextual Information as Input for the Prototype

In the context of the prototypical realization of the developed adaptivity strategy presented above, a simplified approach was chosen with respect to contextual information as external input, such as the driver personality and driving situation. Both types of information are imported as parameters in the form of separately created files, and the contents are fed into the prototype.

5.2.2.1.1 Driving Environment. The driving situation, that is highway (H) or city (C), is encoded in a simplified tabular form (s. Table 5.7) including an anonymized user ID and a driving context ID with code. This means a manual effort on the client side of an experimenter and requires updating the defined parameters according to the current driving conditions. On the prototype side, the last available line is read in each case and thus the specified driving condition is taken into account when selecting a voice prompt. By repeatedly fetching this source before the respective voice output selection, the currently valid driving condition can be taken into account.

Table 5.7: The driving situation, that is highway (H) or city (C), is encoded in a simplified tabular form and fetched by the prototype to account for the currently valid driving condition when selecting a voice output.

User ID	Context ID	Driving context
1	2	H
1	1	C
1	2	H
2	1	C
2	...	

5.2.2.1.2 Driver Personality. Similar to the driving environment, the affiliation of a user to a user cluster is fed into the prototype from an external source. The basis for this is a VBA-based input mask, via which a user is first guided through the questions of the Big Five personality questionnaire (Rammstedt and Danner, 2016; s. usage in Section 5.3). The assessment values are stored in tabular form and are automatically analyzed, so that the user is directly assigned to a user cluster based on his self-assessment, which can be taken into account by the prototype as an input source when selecting a voice prompt. Here, the currently valid cluster is likewise taken into account by repeatedly harvesting the resource prior to selecting voice output.

The assignment to a user cluster is based on the comparison of the user's self-assessment with the identified user clusters as summarized in Table 5.3 (p. 164). A user is thus assigned to a particular cluster if his or her self-assessment matches the centroids of one out of three clusters as closely as possible.

5.2.2.2 Pilot Study: Automatically Identifying Contextual Information

It has been shown in Table 5.4 (p. 166) that based on linguistic features the driving condition can only be identified in dependence of the driver personality: Simple linguistic behavior is not automatically related to driving on a city, and inversely, driving on a highway cannot directly be deduced from a comparably complex language style. In this pilot study, it was therefore investigated whether acoustic features allow for an identification of both contextual factors.

Table 5.8: Overview of extracted acoustic features and their variations. Adapted from Stier *et al.* (2020d, Table 1), licensed under CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`).

Feature	**Description and variations**
Spectral Centroid	mean and standard deviation of spectrum
Energy Difference	difference in energy between low ($<$ 500Hz) and high ($>$ 500Hz) frequency
Intensity, Pitch	maximum, mean, minimum and standard deviation of intensity and pitch
MFCCs and deltas	mel-frequency cepstrum coefficients (16 features) and changes (16 features
Tempo	mean speaking rate (in beats per minute)

Furthermore, the results of this pilot study should indicate an additional approach to automatically identify driver personality and driving situation simultaneously, which may be combined with linguistic features in future research. In the following, the investigation concerning the feasibility to automatically detect contextual information, such as driver personality and driving situation, is presented based on the publication by Stier *et al.* (2020d).

This pilot study is based on the data collected as part of the experiment presented in Chapter 4.1. For this purpose, a data subset of 44 participants with an average of 43 years ($SD =$ 13.24) and a gender distribution of 28 male and 16 female subjects were included for evaluation. Thereby, the data collection comprised 707 answers (per participant: $M =$ 58.97, $SD =$ 17.85), split between 376 for the highway ($M =$ 62.76, $SD =$ 18.67) and 331 for the city ($M =$ 55.17, $SD =$ 16.93). The voice recordings were tailored to the individual user responses ranging from 3 to 400s ($M =$ 41.78, $SD =$ 40.19) without further processing. Based on Landesberger *et al.* (2020), 43 features were extracted (Table 5.8) using `librosa` (McFee *et al.*, 2015) and `Praat` (Boersma and Weenink, 2020). In order to reflect that human personality manifests multiple traits simultaneously, the participants' assessed Big Five personality traits were combined as clustering variables instead of investigating traits individually. Thereby, six user clusters were obtained (SPSS 2-step clustering, av. silhouette 0.4, size ratio 3.25). Additionally, the driving context (DC), during which a response was recorded, was distinguished.

Table 5.9: Distribution of participants and answers on highway and city among identified user clusters. Adapted from Stier *et al.* (2020d, Table 2), licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

UC	**1**	**2**	**3**	**4**	**5**	**6**	$\sum$
Subjects	6	6	8	4	12	8	44
Age	41.7	41.3	45.6	41.8	43.5	41.9	42.9
Openness	3.72	2.90	3.14	4.00	3.75	3.63	3.52
Conscientiousness	3.96	3.85	4.46	4.69	3.94	3.92	4.09
Extraversion	3.56	3.54	3.73	4.75	3.78	3.70	3.78
Agreeableness	3.17	3.53	4.05	3.63	3.81	4.15	3.77
Neuroticism	2.83	2.46	1.89	1.59	2.00	2.75	2.26
Answers H	50	60	65	40	95	66	376
Answers C	47	51	53	37	87	56	331

Note: UC – user cluster; H – highway; C – city.

Table 5.9 summarizes the detailed distribution of participants and their characteristics to the distinguished user clusters. Overall, the clusters comprise between 4 (UC 4) and 12 (UC 5) participants.

A stratified five-fold cross-validation was performed and performance measures (accuracy, precision, recall, f-measure; macro-averaging) were compared for a Multinomial Logistic Regression classifier[7] (MLR), a Random Forest Classifier[8] (RFC) and a Support Vector Machine[9] (SVM). In a first step, user clusters (UC) and driving contexts (DC) were processed individually, before combining them (UC & DC). In either classification scenario, the best result was obtained for RFC with an accuracy < 0.966 and f-measure < 0.965 (Table 5.10). However, all classifiers performed remarkably well with the worst results for MLR with an accuracy $<$ 0.723 and f-measure < 0.715. The application of combined resampling (under-, oversampling, smote) and feature selection (cross-validated recursive elimination) methods did not significantly improve classification results and are thus omitted.

[7] C=80, penalty=l1, solver=liblinear

[8] n_estimators=500, bootstrap=false, max_features=log2

[9] C=0.1, degree=1, gamma=0.001, kernel=linear

Table 5.10: Classification results. Taken from Stier *et al.* (2020d, Table 3), licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

		acc	prec	rec	f
UC	MLR	0.774	0.770	0.762	0.762
	RFC	**0.979**	**0.982**	**0.974**	**0.977**
	SVM	0.815	0.805	0.793	0.793
DC	MLR	0.939	0.939	0.939	0.939
	RFC	**0.990**	**0.990**	**0.991**	**0.990**
	SVM	0.955	0.955	0.956	0.955
UC & DC	MLR	0.723	0.728	0.718	0.715
	RFC	**0.966**	**0.971**	**0.962**	**0.965**
	SVM	0.799	0.798	0.787	0.784

In summary, the results of this pilot study indicate a reliable identification of both the driver personality and the driving situation in HMI as a secondary task. In this context, acoustic features were shown to serve as a suitable source to automatically classify the considered contextual information. They especially led to a precise categorization of the respective driving context, presumably due to the unprocessed driving noises. In addition, the selected acoustic features served as a reliable basis to differentiate between user personality clusters. These results indicate the feasibility of the automated classification of contextual information in dual-task contexts. However, the small size and domain-specific nature of the employed data set represents a clear limitation with regard to the deduction of generally valid conclusions. It will be necessary for future work to validate these findings in the context of real-life driving situations. Based on these observations, future research may also combine acoustic with linguistic features for an enfolding approach to employ the spoken language of a driver to automatically deduce his or her personality and the respective driving condition.

5.3 Evaluation as Real-Life User Study

In this section, the results of a user study are presented validating the deduced strategy for in-vehicle user- and situation-adaptive voice output (s. Section 5.1) by means of a prototipical

implementation (s. Section 5.2). This evaluation study took place as a real-life experiment, where the participants were asked to interact with a dialog system while driving a car in different driving situations.

The following sections provide an overview of the employed methodology, followed by a presentation and discussion of results.

5.3.1 Methodology

This section describes the methodological approach of this user study. In a first step, the participants and experimental design are outlined. Second, the employed material is presented.

5.3.1.1 Participants

A total of 8 German native speakers between 24 and 61 years ($M = 44.50$, $SD = 15.26$) and a gender distribution of 4 male and 4 female subjects participated in this experiment. All of them possessed a valid driver's license.

5.3.1.2 Experimental Design

The goal of this user study was to evaluate the adaptation strategy deduced in this work. For this purpose, the developed prototype with user- and situation-adaptive voice output (s. Section 5.2) was compared with a non-adaptive baseline system. For a realistic evaluation of the adaptation strategy, the study took place in the context of a real-life user study in actual road traffic. The participants were thus asked to interact with two different SDSs while driving in compliance with road traffic regulations under different driving conditions, including a highway and a city (s. Figure 5.6). In that way, each participant experienced four different scenarios.

The choice of the described design allowed to focus on the central elements of this thesis and thus to evaluate the perception of an adaptive SDS in comparison with a non-adaptive baseline system in dependence of an individual user and his or her personality traits, and the respective driving condition.

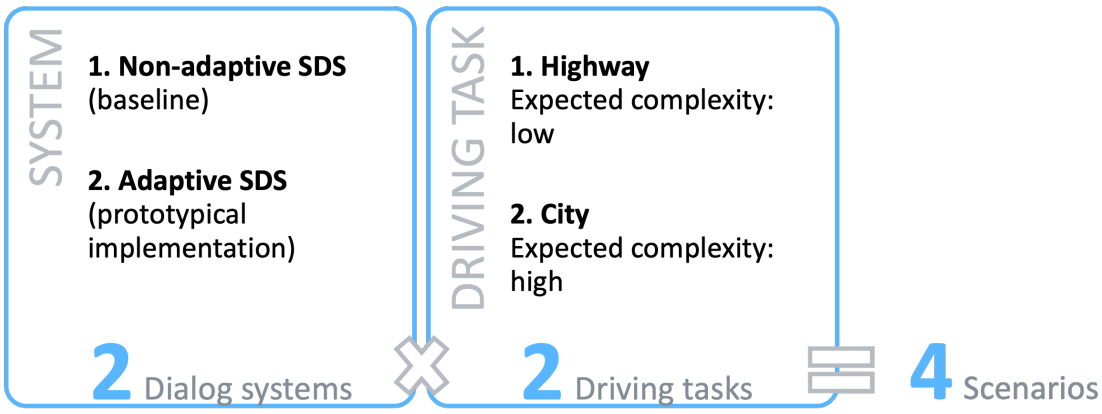


Figure 5.6: The experimental design of the real-life user study focused on the comparison of a baseline with an adaptive SDS in the context of two different driving conditions.

5.3.1.3 Materials

This section provides an overview of the materials of this user study.

5.3.1.3.1 Questionnaires. Two questionnaires were employed to capture demographic and personal data from the participants. In addition, two questionnaires were used to evaluate the perception of the experienced SDSs. All applied questionnaires can be found in Appendix C.3.1 and C.3.2.

- **Preliminary Questionnaire:** Demographic information (age, gender, etc.) about the participants was collected in a pre-survey.
- **Big Five Personality Traits:** The participants were additionally asked to self-assess their personality traits by means of the German version of the BFI questionnaire according to Rammstedt and Danner (2016) on a 5-point scale. It consists of 45 questions assigned to the five personality traits Agreeableness, Conscientiousness, Extraversion, Neuroticism and Openness (s. Section 2.4).
- **DALI:** The DALI questionnaire based on Hofmann (2015) was employed to measure the self-assessed cognitive load of users on the basis of six dimensions (s. Section 2.3) on

a 5-point Likert scale.

– **UEQ:** In addition, the UEQ by Laugwitz *et al.* (2006) was employed to assess the participants' perceived experience with a voice assistant. This questionnaire consists of 26 items in the form of complementary pairs of adjectives assignable to the six factors Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty (s. Section 2.1). The questionnaire was presented in the form of a 7-stage scale.

The preliminary and BFI questionnaires were presented to each participant via a VBA-based interface prior to the user study (Appendix C.3.1), while DALI and UEQ were presented printed in paper format at the half and the end of the experiment.

5.3.1.3.2 Small Talk in Question-Answer Sequences. In order for the prototype to produce adaptive voice output with regards to the individual user and driving situation, small talk questions were employed to gather speech data from the participants, which could be employed by the prototype to define and output a corresponding voice prompt in terms of syntactic complexity. Small talk has been proven as a valuable methodology to record spontaneous, individual language use in a dual-task environment in Chapter 4. In comparison to the user study presented in Section 4.1, where a simulated voice assistant asked questions to be answered by the participants, the present study focused on a real-life dual-task environment. For this reason, the experimenter took on the role of a co-driver, asking questions to be answered to the driver as an interlocutor. The selection of small talk questions was based on simple, personal experiences and private preferences as in the previous user study in order to allow the participants to easily provide answers with reference to their own daily environment. Table 5.11 provides a summary of the included questions and topics (general small talk, leisure time, preferences, travelling).

5.3.1.3.3 Domain, Question & Sentence Types. The domain, question and sentence type have been proven to directly influence the perceived naturalness and comprehensibility of voice prompts in the user studies presented in Section 3.2. In order to account for these factors influencing the perception of an SDS, different domains and question types were considered in this real-life user study for the purpose of user-initiated, one-shot QAS. Similar to the previous studies, the domains DAS and COP were chosen as vehicle-related contexts along with different question types associated with various vehicle functions (s. Table 5.12):

Table 5.11: Small talk topics and questions used in the real-life user study.

Topic	**Example**
General small talk	*Das Wetter ist ja jetzt doch schön winterlich. Wie findest du es?* (eng. "The weather is nice and wintry now. What do you think?")
	Was ist eigentlich deine liebste Jahreszeit, und warum? (eng. "What is actually your favorite season, and why?")
	Was würdest du tun, wenn du morgen im Lotto gewinnen würdest? (eng. "What would you do if you won the lottery tomorrow?")
Leisure time	*Was ist ein Projekt, das du zuhause schon lange einmal umsetzen möchtest?* (eng. "What's a project you've been wanting to do at home for a long time?")
	Welchen Film oder welche Serie hast du zuletzt gesehen und worum ging es darin? (eng. "What was the last movie or series you saw, what was it about?")
	Wie entspannst du am besten nach einem anstrengenden Arbeitstag? (eng. "What's the best way to relax after a hard day at work?")
Preferences	*Hast du eine Lieblingsstadt? Was gefällt dir an ihr so besonders gut?* (eng. "Do you have a favorite city? What do you like so much about it?")
	Wie feierst du gern Silvester? Wie hast du es dieses Jahr verbracht? (eng. "How do you like to celebrate New Year's Eve? How did you spend it this year?")
	Welchen Beruf wolltest du als Kind eigentlich immer haben, und warum? (eng. "What profession did you always want to have as a child, and why?")
Travelling	*Was hältst du von Kreuzfahrten? Welche Erfahrungen hast du gemacht?* (eng. "What do you think about cruises? What experiences do you have?")
	Welche Art von Urlaub machst du lieber: Städtetrip oder Pauschalreise? (eng. "What kind of vacation do you prefer: city trip or all-inclusive?")
	Wo warst du zuletzt im Urlaub? Wie hat es dir gefallen? (eng. "Where were you last on vacation? How did you like it?")

What As defined in Subsection 3.1.2, the conceptual explanation to the question *Was ist Funktion F?* (eng. “What is function F?”) supplies a general definition of a particular function F and focuses on the transfer of factual knowledge.

How The question *Wie funktioniert Funktion F?* (eng. “How does function F work?”) was introduced in Subsection 3.3.1.3.3 and focuses on the transfer of methodological knowledge associated with a concrete function F to describe its functional scope. It thus presupposes the factual knowledge of the respective function. In order to decouple explanations in response to this question type from a preceding conceptual explanation and to allow for one-shot QAS, this question type was only allowed in connection with certain vehicle functions from which the conceptual knowledge can already be inferred intuitively from the name, such as in the case of DAS and the Lane Keeping Assist (*i.e.*, an assistant to maintain the distance to the driver in front).

Which In this user study, an additional question *Welche Funktionen gibt es?* (eng. “Which functions are there?”) was included as a special case of What, asking about the functions that are subsumed under an overarching term.

Overall, four syntactic paraphrases with an increasing complexity were prepared for the vehicle functions associated with the adaptive system (s. Table 5.12). They were manually created following the requirements of explanatory voice output (s. Section 3.1.3) and the methodological approach described in Section 3.1.4. The resulting prompt variants were stored in a `PostgreSQL` database (s. Subsection 5.2.1.5) and are summarized in Appendix C.3.3. Table 5.13 exemplifies the syntactic paraphrases by means of the Attention Assist: While the first complexity level consists of linear main clauses, the second level is characterized by the introduction of an object-oriented relative clause and a coordination of the last main clauses. The third complexity level can be differentiated by a subject-oriented subordination and a pronominalisation of the noun phrase in the second main clause. In comparison to the previous complexity levels, the fourth aggregates to a single main clause, where the object of the first main clause is shifted into a subject-oriented relative clause and reduced to the role of an apposition with, in this case, the conjunction *als* (eng. “as”). With the increasing number of aggregation strategies and paraphrasing of individual constituents, an increasing syntactic complexity is assumed, ranging from the simple, linear structures of main clauses to a nested

Table 5.12: Small talk topics and questions used in the real-life user study.

System	Domain	Question type	Function
Adaptive	COP	What	*Beduftung* (“Perfume Atomizer”)
		How	*Sprach-Assistent* (“Voice Assist”)
		Which	*Massage-Programme* (“Massage Programs”)
	DAS	What	*Aufmerksamkeits-Assistent* (“Attention Assist”)
		How	*Brems-Assistent* (“Brake Assist”)
		How	*Totwinkel-Assistent* (“Blind Spot Assist)
Non-adaptive	COP	What	*4D-Tiefenmassage* (“4D Sound Massage”)
		What	*Lordosen-Stütze* (“Lumbar Pad”)
		Which	*Innovationen* (“Innovations”)
	DAS	How	*Spurhalte-Assistent* (“Lane Keeping Assist”)
		How	*Verkehrszeichen-Assistent* (“Traffic Sign Assist”)
		Which	*Diebstahl-Warnanlage* (“Theft Alarm System”)

syntactic form.

5.3.1.3.4 Prototype and Wizard-of-Oz as Simulated SDS. The prototypical implementation presented in Section 5.2 was used as the adaptive SDS in this study. As a non-adaptive basis for comparison, a Daimler internal tool on the model of SUEDE (Klemmer *et al.*, 2000) was employed in order to simulate spoken interaction between the participants and a real SDS. For this purpose, the dialog flow specification of the user study presented in Section 3.3 was reused: Other than in the previous studies, no dialog task was required to be displayed as a picture on the head-unit. Instead, the experimenter was supposed to ask the participants to formulate a question to the SDS and to additionally request and record a participant’s assessment of the corresponding explanatory prompt. The dialog flows were thus adapted accordingly.

5.3.1.3.5 Experimental Setup. The study was conducted in a Mercedes-Benz A-Class (2019, automatic gearbox) in the area of Ulm, Germany. During the drive, the participant sat in the driver’s seat, while the experimenter in charge sat diagonally behind on the back seat and monitored both the participant and the respective SDS running on a laptop. In order to cap-

Table 5.13: Syntactic paraphrases exemplified for the Attention Assist.

Complexity level	**Syntactic paraphrases**
1	*Der Aufmerksamkeits-Assistent ist ein Sicherheitssystem. Er misst bei langen oder monotonen Fahrten deine Aufmerksamkeit. Der Aufmerksamkeits-Assistent warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit. Gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.* (eng.“The attention assist is a safety system. It measures your attention during long or monotonous journeys. The attention assist warns you with a signal tone when your attention wanes. At the same time, a recommendation to take a break soon appears in the instrument cluster.”)
2	*Der Aufmerksamkeits-Assistent ist ein Sicherheitssystem, das bei langen oder monotonen Fahrten deine Aufmerksamkeit misst. Der Aufmerksamkeits-Assistent warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit und gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.* (eng. “... safety system that measures... when your attention is waning and at the same time...”)
3	*Der Aufmerksamkeits-Assistent, der bei langen oder monotonen Fahrten deine Aufmerksamkeit misst, ist ein Sicherheitssystem. Er warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit und gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.* (eng. “The attention assist, which measures your attention during long or monotonous journeys, is a safety system. It warns you...”)
4	*Der Aufmerksamkeits-Assistent, der als Sicherheitssystem bei langen oder monotonen Fahrten deine Aufmerksamkeit misst, warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit und gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.* (eng. “The attention assistant, which is a safety system that measures your attention during long or monotonous journeys, warns you with a signal tone when your attention is waning and at the same time...”)

ture voice input of the participant, a portable microphone was attached to the subject's seatbelt and connected to the laptop. The device was connected via bluetooth with the in-vehicle sound system to play voice output. The built-in Mercedes-Benz voice assistant was disabled during the experiment.

5.3.1.4 Procedure.

The study was divided into two phases and lasted 45-60 minutes per subject. The following subsections provide a detailed overview of the individual phases.

5.3.1.4.1 Phase 1: Pre-Survey and Instructions. In a first step, each participant was asked to sign a declaration of consent to the collection of personal data and recording of sound material, as well as a non-disclosure agreement. In a second step, the preliminary and BFI questionnaires were presented to each participant via a VBA-based interface (Appendix C.3.1). In that way, the assessed values were stored in an Excel table and automatically fed into the prototype as external contextual information. Subsequently, the content of the study was explained: Each participant was instructed to interact with a voice assistant during a daytime drive on a highway and in a city, without revealing that they would experience two different assistants. The planned route was clarified in detail and the participants were explicitly asked to comply with the applicable road traffic regulations and speed limits, that is 100 km/h on the highway and 50 km/h in the city. They were additionally prepared to anwer small talk questions posed by the experimenter in the role of a co-driver and to ask vehicle-related questions themselved to the SDS in the form of user-initiated QAS. By means of examples, the different domains (COP and DAS) and question types (What, How, Which) were demonstrated (s. Appendix C.3.4), as well as the applied evaluation scale to be considered when assessing the comprehensibility of system prompts (s. Appendix C.3.5). In this context, the concept of comprehensibility was introduced as in the user studies in Chapter 3, in that the participants were instructed to assess on a 5-point Likert scale whether a voice prompt was immediately and intuitively comprehensible without further thought according to their subjective opinion. They were explicitly asked to not consider aspects such as the TTS voice and error-free pronunciation in their ratings.

Before starting the experiment, each participant received an introduction to the vehicle controls.

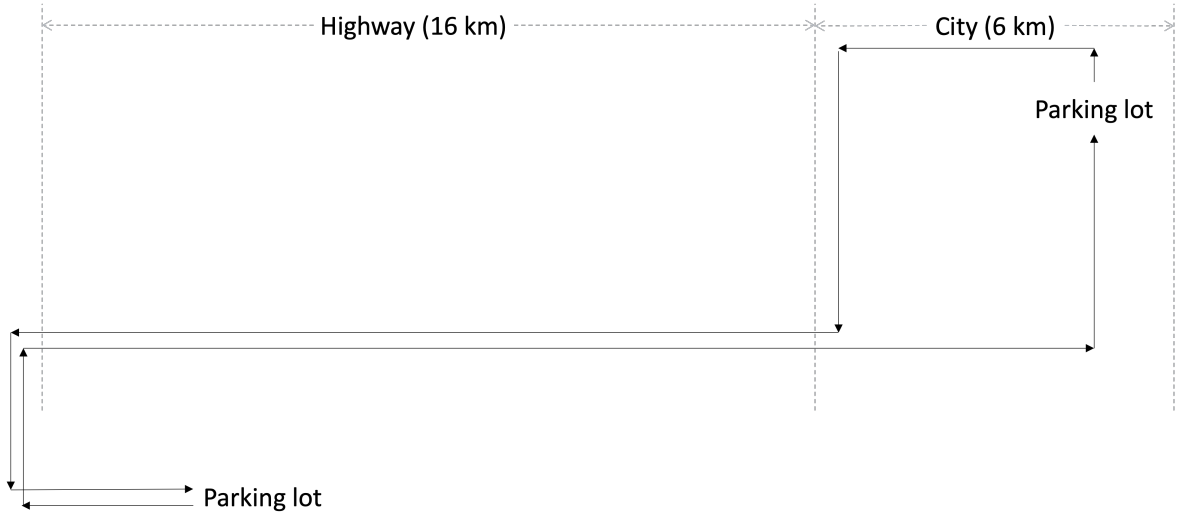


Figure 5.7: The planned route of the real-life user study.

5.3.1.4.2 Phase 2: Experiment. The experiment was carried out in real-life traffic in the area of Ulm, Germany. Starting from a parking lot close to Ulm, the planned route (s. Figure 5.7) led for about 8 kilometers along a straight highway into the city of Ulm. In the city traffic, the itinerary included an approximately 6 km route around the city center with a stop at a parking lot before returning to the highway back to the starting point. In total, the planned route covered about 22 km. The journey of each participant thus took about 30 minutes on average, depending on the traffic volume.

The procedure of the experiment is visualized in Figure 5.8: During the drive, the experimenter assumed the role of a passenger and asked the participant a small talk question, which the participant answered as spontaneously as possible following the previous instructions. The subject was then verbally given the task of asking a vehicle-related question to the voice assistant by the experimenter. To activate the voice assistant, the participants were instructed to speak the phrase *Hallo Mercedes* (eng. “Hello Mercedes”) in order to keep up the illusion of a real SDS interaction. The driver’s question was followed by an explanatory answer from the voice assistant. Following each QAS, the participant was asked by the experimenter to rate the heard answer with respect to its comprehensibility.

The procedure described above was repeated for the randomized small talk questions and

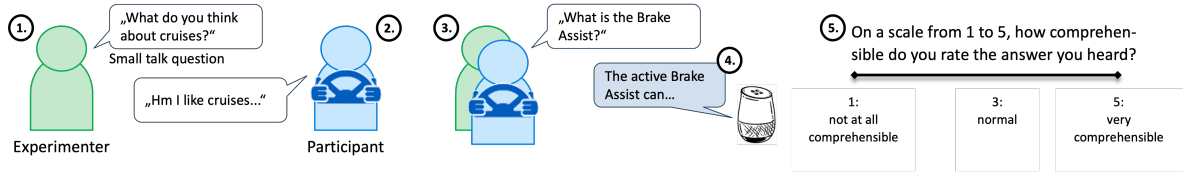


Figure 5.8: Procedure of the real-life user study.

vehicle functions for one of the two systems (*i.e.*, prototype or WoZ). After the first half of the drive, a brief stop was made in a parking lot within the city, during which participants were asked to complete an interim questionnaire to assess the voice assistant they experienced. The described procedure was repeated on the return trip for the second system (*i.e.*, WoZ or prototype) and again concluded with a questionnaire assessing the experienced voice assistant at the starting parking lot. Overall, the order of systems and driving complexities was randomized over participants.

Due to the given circumstances and small number of participants, it was not possible to fully iterate all parameters per system, such as the question type and vehicle function. With the described study design, each participant experienced the same domains and question types associated with the same vehicle functions defined per system. Despite this general limitation, the study design nevertheless allowed to explicitly focus on the perceived difference between the adaptive and non-adaptive systems. It is therefore considered to serve the purpose of the context of this thesis.

5.3.1.5 Dependent Variables

Evaluation measures. Different types of data were collected in the course of this user study, including the personal data collected in the pre- and intermediate/post-survey. Table 5.14 provides an overview of the measures, which were employed in the evaluation based on the collected data. In order to evaluate the perceived difference in experience between the adaptive and non-adaptive systems, the perceived comprehensibility (*Comp*) per voice prompt was recorded by the experimenter during the experiment (6/participant and system). In addition, the *DALI* and *UEQ* questionnaires provided a subjective assessment of the workload and user experience as part of the intermediate and post-survey. The decision to include subjective

Table 5.14: Measures of the real-life user study.

	Measure	Source
User experience	Perceived comprehensibility (*Comp*)	Experimenter logs
	Assessment of user experience (*UEQ*)	UEQ questionnaire (intermediate/ post-survey)
Driver distraction	Assessment of workload (*DALI*)	DALI questionnaire (intermediate/ post-survey

measures to assess both user experience and driver distration is based on prior results on the previous user studies conducted in this work. Thereby, the subjective assessment generally provided valuable insights into the respective research questions, while objective measures, such as driving performance parameters, appeared rather contradicting and did not contribute for the purpose of this work.

Hypotheses. Three hypotheses were formulated on the basis of these measures in order to evaluate the experienced difference between the adaptive system (*ADAPT*) as prototypical implementation of the adaptation strategy developed in this work and a non-adaptive, standard system (*STAND*). They are presented in the following and will be validated with the statistical analyses in the following section.

- The perceived comprehensibility of voice prompts was shown to be influenced by their syntactic form (s. Section 3.3). Based on this observation, a user-and situation-adaptive strategy for voice output was deduced by relating syntactic preferences with syntactic characteristics in driver language and realized in the context of a prototypical implementation. By taking external contextual factors such as user personality and driving condition into account, the adaptive system *ADAPT* is thus assumed to be perceived as more comprehensible while driving compared to the non-adaptive standard system *STAND*.

$$Comp_{\text{STAND}} < Comp_{\text{ADAPT}} \tag{5.2}$$

- The avowed objective of the adaptivity strategy developed in this work is to provide an improved user experience compared to a standard system that does not vary in its voice output according to external contextual factors. Accordingly, the adaptive system *ADAPT* is assumed to be perceived with a better user experience compared to the

standard system *STAND*. This specifically applies to the hedonic quality aspects Stimulation and Novelty, while pragmatic quality aspects such as Efficiency, Dependability and Perspicuity are not primarily focused in this context as both systems follow the same goal-oriented QAS approach.

$$UEQ_{\text{STAND}} < UEQ_{\text{ADAPT}} \tag{5.3}$$

- Driving as primary task induces a certain cognitive load on the driver. When interacting with an SDS as secondary task while driving, the cognitive load is generally known to increase. In this context, the cognitive load induced by a non-adaptive standard system *STAND* is expected to be higher compared to our adaptive system *ADAPT*, which takes external contextual factors into account and generates voice output specific to the needs of a user and situation.

$$DALI_{\text{ADAPT}} < DALI_{\text{STAND}} \tag{5.4}$$

The outlined real-life user study was conducted according to the presented methodology. In the following sections, the results of the experiment are presented followed by a discussion and validation of the above hypotheses.

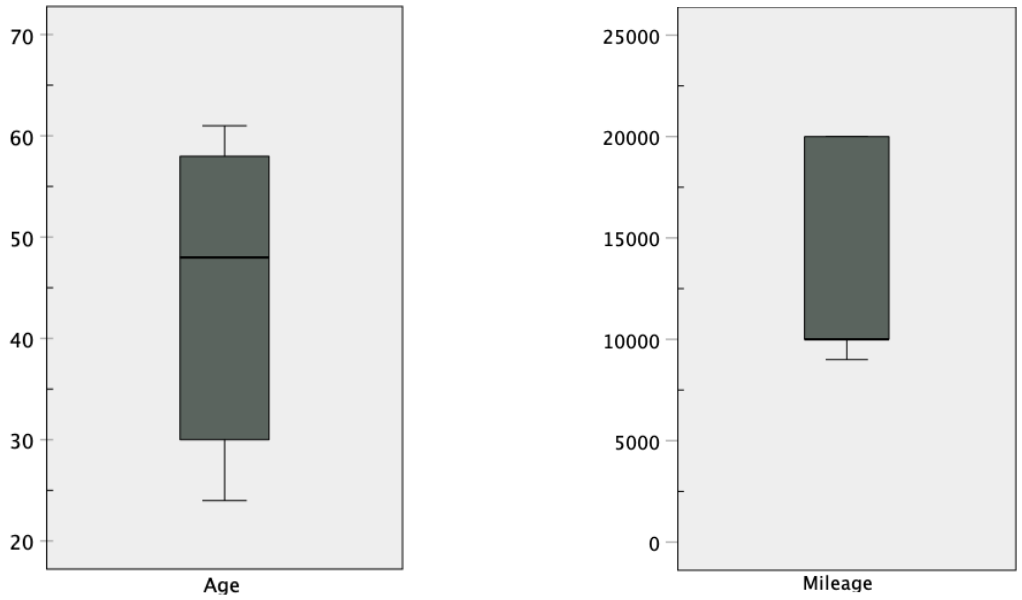
5.3.2 Statistical Analyses and Results

In this section, the results of the statistical analyses are presented. First, the results of the pre-survey will be described, followed by subjective assessment of the user experience and cognitive workload. Asterisks are employed to indicate if and at which level a comparison of conditions was found to be statistically significant (* $p < .05$, ** $p < .01$, *** $p < .001$).

5.3.2.1 Questionnaire Results

Demographic and personal information of the participants was collected prior to the experiment. Overall, the data of 6 participants was included in analyses with a gender distribution of 3 male and 3 female subjects.[10] The average age was 44.83 years ($SD = 15.54$) within a range from 24 to 61 years ($Mdn = 48$ years; s. Figure 5.9a). The average annual mileage was indicated as 13,166.67 km ($SD = 5{,}307.23$) ranging between 9,000 and 20,000 km per year ($Mdn = 10{,}000$ km; s. Figure 5.9b).

[10] Two participants were excluded from analysis due to technical problems during the experiment.



(a) Box plot regarding age.

(b) Box plot regarding annual mileage.

Figure 5.9: Results of the real-life user study concerning the age and annual mileage (median as vertical band in the box center).

Figure 5.10 summarized the individual Big Five personality traits according to Rammstedt and Danner (2016) on a 5-point Likert scale. Overall, the participants indicated homogeneously to be tolerable (Agreeableness: $M = 3.72$, $Mdn = 3.70$, $SD = 0.26$, $IQR = 0.28$) and on average rather conservative towards new experiences (Openness: $M = 2.82$, $Mdn = 3.00$, $SD = 0.67$, $IQR = 0.65$). They assessed themselves as generally orderly and disciplined (Conscientiousness: $M = 3.65$, $Mdn = 3.61$, $SD = 0.69$, $IQR = 0.97$) and average extraverted (Extraversion: $M = 3.17$, $Mdn = 3.31$, $SD = 0.66$, $IQR = 1.13$). Although the participants considered themselves on average to be rather mildly neurotic (Neuroticism: $M = 2.67$, $Mdn = 2.88$), the greatest variance among the personality traits was evident in the self-assessment regarding neuroticism ($SD = 0.85$, $IQR = 1.13$).

Based on their self-assessed Big Five personality dimensions, the participants were automatically assigned to one out of three user clusters (s. Subsection 5.2.2.1). Three subjects were assigned to UC 3, while two participants were assigned to UC 2 and 1 subject was assigned to UC 1. The distribution of their characteristics is listed in Appendix C.3.6 for com-

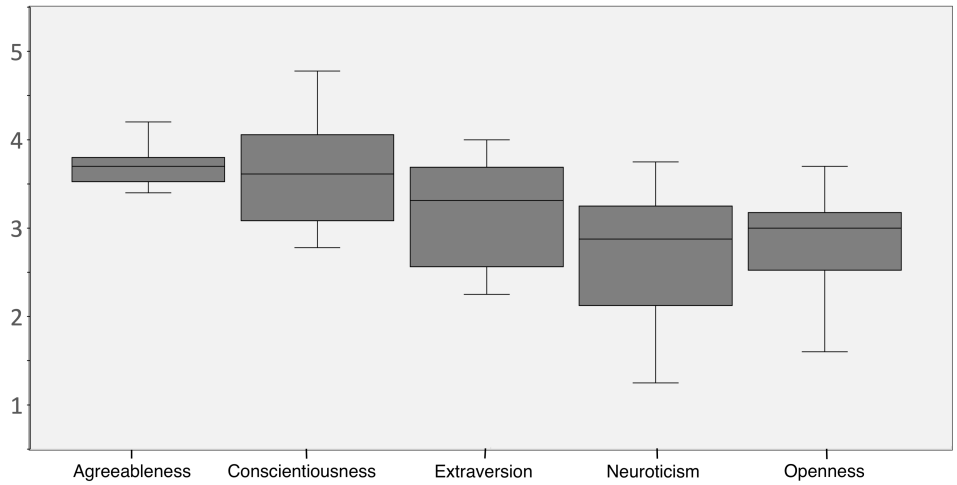


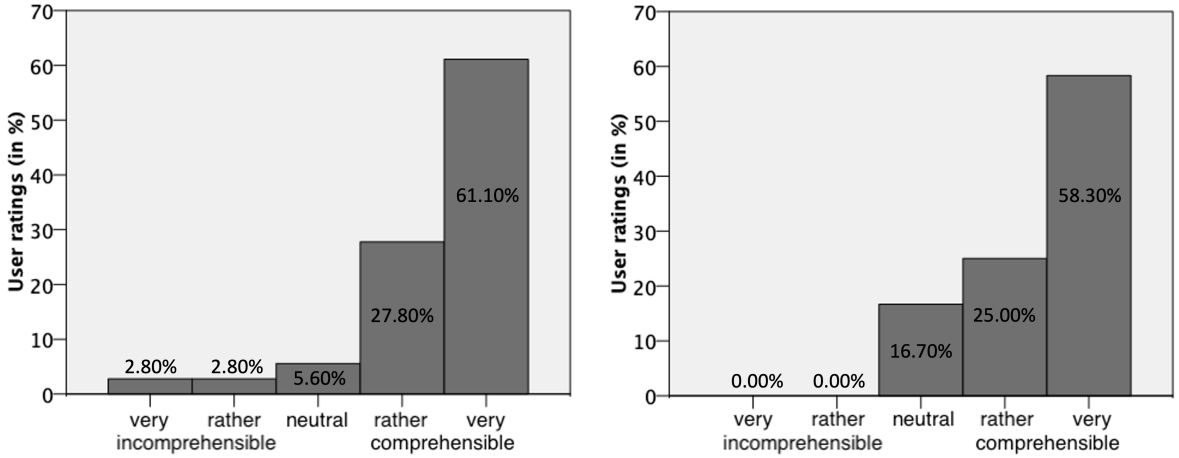
Figure 5.10: The individual Big Five personality traits according to Rammstedt and Danner (2016) on a 5-point Likert scale.

pleteness.

5.3.2.2 Assessment of User Experience

The assessment of user experience was achieved by employing two different evaluation measures. As a first one, the participants' ratings of the perceived comprehensibility of voice prompts was recorded on a 5-point Likert scale. Second, the UEQ questionnaire (Laugwitz *et al.*, 2006) on a 7-point Likert scale. The participants were asked to complete this questionnaire twice, after the first and return trip, for the different systems they experienced.

5.3.2.2.1 Assessment of Comprehensibility. The participants of this user study were asked to assess the perceived comprehensibility of explanatory system answers repeatedly. Overall, 72 voice prompts (12 per participant) were assessed, that is six ratings per user and system. A summary of the assessments is visualized in Figure 5.11. The voice prompts of both systems were generally rated as very comprehensible with a mean value $M = 4.42$ (Mdn= 5.00, $SD = 0.94$) for *ADAPT* and likewise a mean value $M = 4.42$ (Mdn= 5.00, $SD = 0.77$) for *STAND*. On this basis, no statistically significant difference was observed. For completeness,



(a) Comprehensibility ratings for *ADAPT*.

(b) Comprehensibility ratings for *STAND*.

Figure 5.11: Summary of assessments concerning the perceived comprehensibility of voice prompts.

Appendix C.3.7 provides an overview of the distribution of user ratings and syntactic variants per user cluster.

Table 5.15 provides a summary of the perceived comprehensibility based on user cluster and driving condition. The direct comparison between *ADAPT* and *STAND* revealed minor differences in the ratings of the respective voice output. While *STAND* tended to receive better ratings for UC 3 with regard to the comprehensibility of prompts (C: $M = 4.67$, $Mdn = 5.00$, $SD = 0.71$, $IQR = 0.00$; H: $M = 4.33$, $Mdn = 4.00$, $SD = 0.71$, $IQR = 1.00$) than *ADAPT* (C: $M = 4.22$, $Mdn = 4.00$, $SD = 0.83$, $IQR = 1.00$; H: $M = 4.22$, $Mdn = 4.00$, $SD = 0.97$, $IQR = 1.00$), this was mainly the case for UC 2 in the city (*ADAPT*: $M = 4.67$, $Mdn = 5.00$, $SD = 0.52$, $IQR = 1.00$; *STAND*: $M = 4.83$, $Mdn = 5.00$, $SD = 0.41$, $IQR = 0.00$). In contrast, *ADAPT*'s voice output for UC 2 was clearly rated better on the highway (*ADAPT*: $M = 4.83$, $Mdn = 5.00$, $SD = 0.41$, $IQR = 0.00$; *STAND*: $M = 4.00$, $Mdn = 4.00$, $SD = 0.89$, $IQR = 2.00$). This observation holds in reverse for UC 1, in that voice prompts were indicated as more comprehensible for *STAND* on the highway (*ADAPT*: $M = 3.67$, $Mdn = 5.00$, $SD = 2.31$, $IQR = 4.00$; *STAND*: $M = 4.00$, $Mdn = 4.00$, $SD = 1.00$, $IQR = 2.00$) and more comprehensible for *ADAPT* in the city (*ADAPT*: $M = 5.00$, $Mdn = 5.00$, $SD = 0.00$, $IQR = 0.00$; *STAND*: $M = 4.33$, $Mdn = 5.00$, $SD = 1.15$, $IQR = 2.00$). However, except for

Table 5.15: The perceived comprehensibility per user cluster (UC) and driving condition (DC).

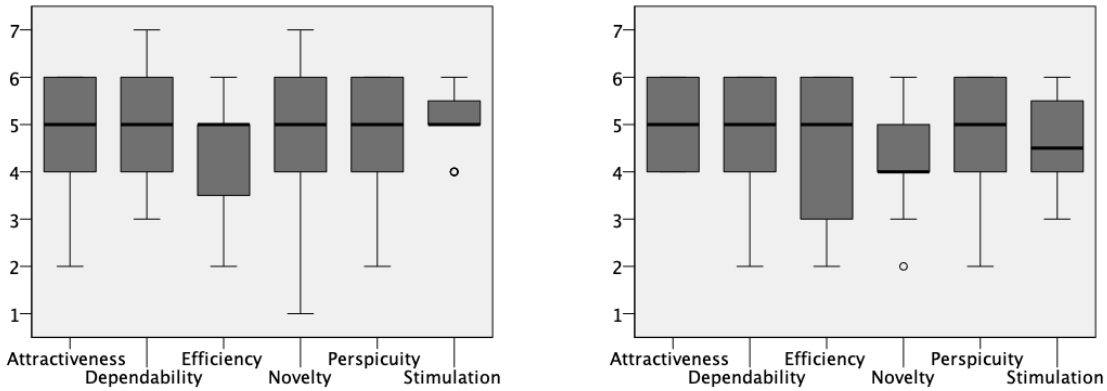
System	UC	DC	*COMP*			
			M	Mdn	SD	IQR
ADAPT	**1**	**C**	5.00	5.00	0.00	0.00
		H	3.67	5.00	2.31	4.00
	2	**C**	4.67	5.00	0.52	1.00
		H	4.83	5.00	0.41	0.00
	3	**C**	4.22	4.00	0.83	1.00
		H	4.22	4.00	0.97	1.00
STAND	**1**	**C**	4.33	5.00	1.15	2.00
		H	4.00	4.00	1.00	2.00
	2	**C**	4.83	5.00	0.41	0.00
		H	4.00	4.00	0.89	2.00
	3	**C**	4.67	5.00	0.71	0.00
		H	4.33	4.00	0.71	1.00

Note: UC – User Cluster; DC – Driving condition.

these general trends, no statistically significant difference was observed.

In these regards, the above observations are contrary to Hypothesis 5.2.

5.3.2.2.2 User Experience Questionnaire. The results of the six UEQ factors are summarized in Figure 5.12. Overall, the ratings of both systems appear quite similar with only slight differences. As such, both systems were perceived as similarly attractive (Attractiveness; *ADAPT*: $M = 4.80$, $Mdn = 5.00$, $SD = 1.06$, $IQR = 2.00$; *STAND*: $M = 5.00$, $Mdn = 5.00$, $SD = 0.79$, $IQR = 2.00$) and predictable (Dependability; *ADAPT*: $M = 5.00$, $Mdn = 5.00$, $SD = 1.03$, $IQR = 2.00$; *STAND*: $M = 4.90$, $Mdn = 5.00$, $SD = 1.12$, $IQR = 2.00$). Similarly, the participants rated both systems as efficient (Efficiency; *ADAPT*: $M = 4.35$, $Mdn = 5.00$, $SD = 1.23$, $IQR = 1.75$; *STAND*: $M = 4.50$, $Mdn = 5.00$, $SD = 1.40$, $IQR = 3.00$) and transparent (Perspicuity; *ADAPT*: $M = 4.60$, $Mdn = 5.00$, $SD = 1.23$, $IQR = 2.00$; *STAND*: $M = 4.70$, $Mdn = 5.00$, $SD = 1.46$, $IQR = 2.00$). A clear difference was perceived concerning the factors Stimulation and Novelty. According to the participants' ratings, *ADAPT* conveyed a statistically significant more novel character ($M = 4.80$, $Mdn = 5.00$, $SD = 1.32$, $IQR =$



(a) UEQ factors for *ADAPT*.

(b) UEQ factors for *STAND*.

Figure 5.12: The UEQ factors according to Laugwitz *et al.* (2006) on a 7-point Likert scale.

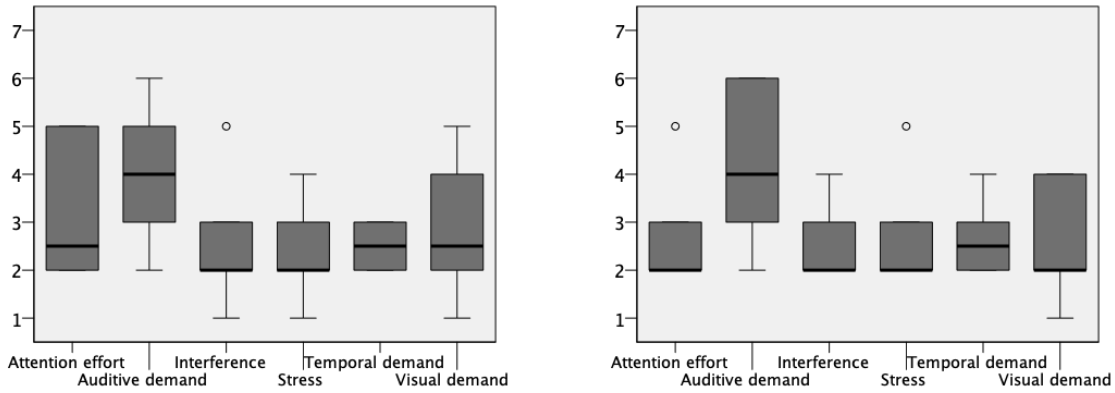
2.00) than *STAND* ($M = 4.25$, $Mdn = 4.00$, $SD = 1.12$, $IQR = 1.00$; $Z =$ -2.392, $p < .05$, $r = .38$). Likewise, compared to *STAND* ($M = 4.55$, $Mdn = 4.50$, $SD = 1.10$, $IQR = 1.75$), *ADAPT* was perceived as more stimulating ($M = 5.05$, $Mdn = 5.00$, $SD = 0.69$, $IQR = 0.75$; $Z =$ -2.153, $p < .05$, $r = .34$).

Overall, the above observations partially confirm Hypothesis 5.3. Although the participants did not indicate a clear difference in user experience in four out of six UEQ factors, *ADAPT* clearly obtained higher ratings in terms of the perceived Stimulation and Novelty of the system compared with *STAND*.

5.3.2.3 Assessment of Cognitive Workload

The DALI questionnaire based on Hofmann (2015) was employed for the subjective assessment of cognitive workload. It was recorded for both systems separately on a 5-point Likert scale, that is after the first and return trip.

The results of the individual DALI dimensions are visualized in Figure 5.13. Overall, both systems were assessed to induce a similar cognitive load (*ADAPT*: $M = 2.89$, $Mdn = 2.50$, $SD = 1.30$, $IQR = 2.00$; *STAND*: $M = 2.86$, $Mdn = 2.00$, $SD = 1.27$, $IQR = 1.75$) without



(a) DALI dimensions for *ADAPT*.

(b) DALI dimensions for *STAND*.

Figure 5.13: The DALI dimensions based on Hofmann (2015) on a 7-point Likert scale.

a statistically significant difference. This observation is mirrored by the individual dimensions: On average, a low Attentional effort was estimated for *ADAPT* ($M = 3.17$, $Mdn = 2.50$, $SD = 1.47$, $IQR = 3.00$) and *STAND* ($M = 2.67$, $Mdn = 2.00$, $SD = 1.21$, $IQR = 1.50$). Similarly, both systems were assessed with low ratings for the dimensions Interference (*ADAPT*: $M = 2.50$, $Mdn = 2.00$, $SD = 1.38$, $IQR = 1.75$; *STAND*: $M = 2.50$, $Mdn = 2.00$, $SD = 0.84$, $IQR = 1.25$) and Stress (*ADAPT*: $M = 2.33$, $Mdn = 2.00$, $SD = 1.03$, $IQR = 1.50$; *STAND*: $M = 2.67$, $Mdn = 2.00$, $SD = 1.21$, $IQR = 1.50$). Likewise, the Temporal (*ADAPT*: $M = 2.50$, $Mdn = 2.50$, $SD = 0.55$, $IQR = 1.00$; *STAND*: $M = 2.67$, $Mdn = 2.50$, $SD = 0.67$, $IQR = 1.25$) and Visual demands (*ADAPT*: $M = 2.83$, $Mdn = 2.50$, $SD = 1.47$, $IQR = 2.50$; *STAND*: $M = 2.50$, $Mdn = 2.00$, $SD = 1.23$, $IQR = 2.25$) were estimated as generally low. Although not statistically significant, the Auditive demands rating was prominent among the other DALI dimensions. Compared with the generally low-rated workload of the other dimensions, the auditory load induced by *ADAPT* ($M = 4.00$, $Mdn = 4.00$, $SD = 1.41$, $IQR = 2.50$) and *STAND* ($M = 4.17$, $Mdn = 4.00$, $SD = 1.72$, $IQR = 3.25$) was rated comparatively high.

Overall, the above observations are contrary to Hypothesis 5.4.

Table 5.16: Validation of Hypotheses.

$Comp_{STAND} < Comp_{ADAPT}$	✗	(5.2)
$UEQ_{STAND} < UEQ_{ADAPT}$	(✓)	(5.3)
$DALI_{ADAPT} < DALI_{STAND}$	✗	(5.4)

5.3.3 Discussion of Results

This section provides a summary and discussion of the results observed in the real-life user study, which has been conducted to evaluate the adaptation strategy deduced and presented in this chapter. For this purpose, an overview of the respective hypotheses and their validation is provided in Table 5.16.

- Based on the observations in Section 3.3 concerning the strong correlation of the concepts naturalness and comprehensibility, this user study focused on the assessment of the perceived comprehensibility of voice output. Contrary to expectations, the results in a direct comparison of *ADAPT* with *STAND* do not reveal a clear difference in the perceived comprehensibility of their voice prompts. This indicates that the standard voice prompts were rated similarly comprehensible as the adaptive ones and that the adaptation strategy does not add benefit with respect to the comprehensibility of voice prompts. In this respect, Hypothesis 5.2 has to be rejected as the adaptation strategy does not meet the objective to increase the perceived comprehensibility of voice prompts in the dual-task environment of driving. Nevertheless, based on the ratings, it can be summarized that the voice output of both systems were perceived as generally very understandable. This implies that the adaptation of syntactic complexity depending on the user and driving situation was not considered as a complicating factor. In short, the adaptation strategy neither contributed to an increase nor to a reduction of the perceived comprehensibility of voice output. Due to the low number of participants in this experiment, especially when analyzing the results in the individual user clusters in more detail, these observations can only serve as tendencies that would be valuable to investigate in more depth in further research.
- The adaptation strategy presented in this chapter was in short deduced by relating syntactic preferences with syntactic characteristics in driver language. By taking ex-

ternal contextual factors into account, *ADAPT* generates voice output that specifically addresses a driver's needs in terms of syntactic preferences as a reflection of his or her personality and driving situation, and is thus intended to provide an improved user experience compared with *STAND*. The analysis of UEQ partially confirms this expectation. Indeed, no statistically meaningful difference was identified between *ADAPT* and *STAND* with respect to which of the two systems would be characterized as more efficient, transparent or predictable. In terms of these pragmatic quality factors focusing on goal- or task-oriented aspects, this observation was expected and anticipated in Section 5.3.1.5 since both systems *ADAPT* and *STAND* follow the same goal-oriented QAS approach. Nevertheless, the results demonstrate a clearly preferable view of *ADAPT* with respect to the UEQ dimensions Stimulation and Novelty as hedonic quality aspects. In this sense, *ADAPT* surpasses the stimulating effect and originality of *STAND* and its user experience increases beyond mere usefulness. Finally, in terms of the assessed Attractiveness as a global (dis)approval of the two systems, there was no apparent difference between the highly rated *ADAPT* and *STAND*. According to Laugwitz *et al.* (2006), the assessment of Attractiveness results from a weighted evaluation of the individual pragmatic and hedonic quality aspects and can therefore be explained by the above observations. Overall, these results partially confirm Hypothesis 5.3. Nevertheless, they should be interpreted with caution due to the low number of participants.

- The user- and situation-specific voice output of *ADAPT* according to the developed adaptation strategy was designed to reduce the cognitive load of the user induced by the interaction with an SDS while driving compared to *STAND*. However, in a direct comparison of *ADAPT* and *STAND*, no statistically significant differences were identified. Based on the results, at most tendencies can be revealed, which should, however, be interpreted with caution due to the low number of participants in this study. At this point, the Interference factor should be emphasized, which, as an interaction of the different cognitive load sources, was on average equally low for both systems with 2.50 on a 7-point Likert scale, albeit with a higher variance for *ADAPT*. Overall, cognitive load in terms of the individual DALI dimensions was found to be rather low for both systems, except for auditory demands with a slightly higher than average load. These observations contradict Hypothesis 5.4 in that the adaptation strategy does not serve the objective of reducing cognitive load in the interaction with *ADAPT* compared to *STAND*. However, the adaptation of syntactic complexity was assessed to induce a similar level of cognitive load as the standard prompts and in fact was not considered to increase workload.

In this regard, the application of the adaptation strategy shows the same effect as its absence.

In the above validation of the initially formulated hypotheses, it became apparent that two out of three were rejected based on the study results. It was even revealed that the application of the adaptation strategy developed in this work performs the same as its absence in terms of the perceived comprehensibility of voice output (s. Hypthesis 5.2) and the cognitive load induced by HMI as secondary task (s. Hypothesis 5.4). At this point, the hypotheses discussed above should therefore be interpreted in context: Besides the derived implications based on the study results for Hypthesis 5.2 and Hypthesis 5.4, especially the results on user experience in the context of Hypthesis 5.3 represent the determining factor in the positive or negative evaluation of the developed adaptation strategy. Although *ADAPT* as a prototypical realization of the adaptation strategy does not differ from *STAND* in terms of its Attractiveness and pragmatic quality aspects such as Efficiency, Dependability, and Perspicuity due to their common goal-oriented QAS approach, hedonic quality characteristics are in the spotlight in the comparison of *ADAPT* and *STAND*. As described above, the study results showed that *ADAPT*, and thus the adaptation strategy, enhanced the user experience in the UEQ dimensions of Stimulation and Novelty compared to *STAND*. Although the intended effect of increasing the comprehensibility of voice output and reducing cognitive load by syntactically adapting voice prompts to external contextual factors cannot be demonstrated, the application of the adaptation strategy nevertheless led to an enhanced user experience for the study participants. In these regards, the working hypotheses underlying the present research work are affected. As such, Working Hypothesis 1 is generally supported by the findings of this user study. In contrast, Working Hypothesis 2 could not be confirmed on the basis of the subjectively assessed study results. Although a general effect of syntactic forms in voice output on cognitive load and resulting driving performance has been shown (s. Section 3.3), the subjective participant ratings does not support this observation in the evaluation of *ADAPT*. It is therefore conceivable that the influence of syntactic differences is not sufficiently reflected in the adaptation strategy, either in terms of the user or driving context. Objective assessment methods could contribute to validate this assumption, but will need to be verified in future research. In addition, the adaption strategy can be refined, for instance, by focusing on a more detailed user model including age as a potential influencing factor identified in Section 3.3.

At this point, it should again be explicitly pointed out that the results described in this section are to be interpreted taking into account certain limitations. This concerns first of all the

low number of participants and, thereby, the choice of the study design and scope. It is beyond discussion that the limited amount of data collected has only limited statistically reliable significance. For the evaluation of two systems depending on two driving conditions and three user clusters, it is assumed that at minimum the duplication of participants would be required to achieve this. Nevertheless, the user study described in this section represents the available options at the time of execution and may be considered as a basis for future research.

5.4 Summary of Results

This section provides a summary of the development, realization and evaluation of the adaptation strategy concerning user- and situation-adaptive voice output in dual-task environments, which has been presented in the course of this chapter. First, the procedure of development and realization are summarized. Second, the evaluation study and most relevant results are recapped. Finally, implications on following research are outlined.

In this chapter, the development procedure and deduction of a user- and situation-adaptive strategy with a focus on the syntactic complexity of SDS voice output in the dual-task environment of driving was described. For this purpose, the aspects of language perception (Chapter 3) and production (Chapter 4) were combined. More precisely, the findings concerning syntactic preferences of in-vehicle voice output were compared with the syntactic characteristics of driver speech under consideration of a driver's personality and the driving condition, which have been identified as factors influencing both the perception and production of spoken language. The proposed adaptation strategy (s. Table 5.3, p. 5.3) takes three different user clusters according to the Big Five personality traits (UC 1: the "extraverts"; UC 2: the "neurotics"; UC 3: the "conscientious") and two driving conditions with varying complexity (city C: complex task; highway H: simple task) into account. Following this procedure led to the conclusion that the application of one single adaptation principle in HMI as secondary task cannot be considered as a universally valid strategy. In fact, the choice to either mirror or contrast a driver's behavior depends on both individual user and situation characteristics. As such, the deduced adaptation strategy specifies a contrasting behavior for UC 1 and a mirroring behavior for UC 2, while UC 3 requires a contrasting and mirroring behavior on a highway and in a city, respectively. In this regard, the presented research demonstrated that in the context of a dual-task environment like driving, both the *similarity attraction* (*e.g.*, Nass *et al.*, 1995), where users prefer an SDS whose voice output reflects their own linguistic behavior and

thereby manifests a similar personality, and the *complementarity principle* (*e.g.*, Isbister and Nass, 2000), where a user prefers an opposite personality with deviating linguistic behavior, are applied in dependence of the driving condition.

The deduced adaptation strategy was realized by means of a prototipical implementation within the framework of a JavaScript websocket application (s. Figure 5.4, p. 170), whose detailes are described in the second section of this chapter. The application is focused on the domain of Question-Answer Sequences and follows the generally known SDS architecture with an extension of the Dialog Manager. In summary, it contains a Syntactic Analysis Module to analyze an ASR transcript and to compute its syntactic complexity level, which is fetched by the Disambiguation Module to translate it, under consideration of additional contextual information such as the user personality and driving condition, into a targeted voice output complexity level according to the defined adaptation strategy. The definition of a voice prompt consists of a database query and the retrieval of one out of a set of syntactic paraphrases performed by the Natural Language Generation Module. A pilot experiment was conducted to investigate the reliability to compute syntactic complexity scores on the basis of raw ASR strings. Here, the comparison of syntactic complexity levels computed at runtime in a driving simulation setup and offline employing manual annotations indicated an accuracy of 68.93%, which was concluded as sufficiently reliable to allow for further evaluations of the adaptation strategy. Subsequently, it has been introduced that contextual information in the context of the prototipical realization is encoded in simplified tabular form and used as input by the prototype. Given this known limitation, an additional pilot study was conducted to demonstrate the feasibility to automatically identify contextual information, that is user personality and driving condition.

Finally, a real-life user study was performed to evaluate the deduced adaptation strategy by means of the prototipical realization. The study design involved the participants interacting with the prototype providing user- and situation-adaptive voice output and a non-adaptive WoZ while driving under different conditions. As such, the drive took place in actual road traffic with one part on a highway and one in a city. Besides driving in compliance with the road traffic regulations, the participants were supposed to answer small talk questions from the experimenter in the role of a co-driver. The thereby produced linguistic material served as input for the prototype, which in turn provided explanatory voice prompts of a particular syntactic complexity as answers to vehicle-related, user-initiated QAS. While the output of the non-adaptive baseline system was in the form of linear main clauses, the prototype employed syntactic

paraphrases on a complexity continuum between the poles MCV and RCV. Contrary to expectations, the results of this experiment indicated that the application of the adaptation strategy did not reveal an improving or deteriorating effect concerning the perceived comprehensibility of voice prompts nor the cognitive load induced by HMI as secondary task compared to the non-adaptive system. In this regard, the Working Hypothesis 2 underlying the present research work could not be comfirmed. However, it has been shown that the adaptation strategy enhanced the user experience in terms of the hedonic quality aspects Stimulation and Novelty. In general, Working Hypothesis 1 can therefore be confirmed. As indicated, the results of this study should be interpreted with caution due to the known limitations.

Overall, the adaptation strategy presented in this chapter, which focuses on the syntactic design of voice output in SDS interaction as a secondary task, relies on experience and usability from a user perspective. Its development was underpinned with various resources and supported with several pilot studies. In this context, a number of limitations was identified. Nonetheless, it is considered a valuable contribution and foundation for future research.

Chapter 6

Conclusion and Future Directions

Voice assistants have been entering everyday life for quite a long time now. Be it in the form of Amazon's Alexa, which reads out the recipe while cooking, to Mercedes' MBUX in the car, which supports a driver by simplifying the selection of a route to the excursion destination. What unifies the voice assistants in the above examples is the voice-based interaction with a user in a so-called dual-task environment alongside a focused primary task and completing a task as efficiently as possible. What differentiates them, however, is the context of usage: Compared to the interaction with Alexa, there is a comparatively greater safety aspect associated with the interaction in the vehicle, because the voice-based interaction must not distract the driver from the primary task of driving and instead has to proceed in parallel. While a user will automatically adapt his or her linguistic behavior to the external influences caused by the driving context, the design of the voice output of an SDS, among others, is one factor that is directly perceptible by the user, that is whether communication as a secondary task has a disruptive or supportive effect on the primary task. The requirements for the voice output of a voice assistant in a vehicle are therefore to minimize driver distraction while providing the most positive user experience possible. Even the syntactic design of voice prompts can have an influence in this regard. The adaptive design of SDSs can be supportive in this context in order to address the immediate needs of a driver and the respective driving situation. The thesis at hand addressed the topic of user- and situation-adaptive voice output in SDS interaction as a secondary task with a focus on syntactic complexity.

6.1 Overall Summary and Research Contributions

The goal of this thesis was the development, realization, and evaluation of an adaptive strategy for dialog systems in dual-task environments such as driving, which adapts its voice output with respect to its syntactic complexity while taking the individual user and driving condition into account. In this context, this research work was strictly focused on the scenario of driving. Since a natural language SDS can support numerous different domains, the focus of the present work was additionally limited to one-shot Question-Answer Sequences and the context of vehicle-related functions. For the purpose of deducing an adaptation strategy under consideration of the user perspective and user experience, several user studies were conducted on the basis of underpinning foundational work to define a theoretical framework. Overall, the following theoretical, practical and experimental contributions were described in the research work in the course of this thesis.

The research goal of this thesis required the provision of various fundamentals as, to the best of our knowledge, no comparable literature or approaches exist. Thus, on a theoretical level, requirements for explanatory voice prompts according to Grice (1975)'s maxims were defined in a first instance. On this basis, a procedure for the manual preparation of syntactic paraphrases was defined in order to enable the controlled investigation of the influence of syntactic forms in vehicular voice output. In this context, the approach to collect data on user preferences via the audio channel was validated. Based on this basic research, it was then possible to investigate the role of syntactic forms in language perception and production. Hereby, the relevance of syntactic forms and their inherent complexity in the perception of voice output in SDS interaction as a secondary task was demonstrated besides the identification of various user- and system-dependent influencing factors. The publications related to these theoretical contributions are Stier and Sigloch (2019) and Stier *et al.* (2020b). Furthermore, in the area of language production as a secondary task, the identification of syntactic complexity features in spoken language enabled a syntax-related characterization of drivers' linguistic behavior in the vehicle. Here, the focus of analyses was on speech behavior in dependence of both a framework of individual user personality traits and the complexity of the primary task conditioned by the interaction context driving. Related publications are Stier *et al.* (2020c,e). Finally, on the basis of these theoretical findings, an approach to deduce an interaction strategy was proposed by combining the aspects of language perception and production. Following this, a user- and situation-adaptive strategy for voice output with respect to its syn-

tactic form and complexity was developed. In addition, the reliability of estimating syntactic complexity and the feasibility to automatically identify contextual information was confirmed on a theoretical level. The publication related to this contribution is Stier *et al.* (2020a,d).

In the course of this work, the theoretically elaborated contributions were enabled in the context of practical applications. Accordingly, the practical part of this work comprises, on the one hand, the provision of the necessary basics, ranging from the preparation of syntactic paraphrases to the validation to assess user preferences from auditive input. Furthermore, several pilot and user studies were planned and conducted in order to validate the aforementioned fundamental approaches and, against the background of this research work, to exploratively investigate the influence of syntactic forms on speech perception and production. In this context, particularly a large-scale driving simulation study to create a corpus of spontaneous spoken driver language, based on manual transcriptions by means of a proprietary guideline, should be emphasized. Related publications for these practical contributions are Stier and Sigloch (2019) and Stier *et al.* (2020b,c,e). Based on the extensively performed syntactic analyses and assessed user preferences concerning syntactic forms in voice output, a dialog strategy for syntactic adaptivity was designed and implemented in the context of a prototypical realization. As a further practical contribution of this research work, the developed strategy was evaluated, and supported by two additional pilot studies to prove the reliability of spoken language input to estimate syntactic complexity and identify driver personality and context. The publications related to these contributions are Stier *et al.* (2020a,d).

On the experimental level, numerous pilot and user studies were conducted in the driving simulator and real vehicle for achieving the research goal of this work. Among others, this included the validation of the underlying principles of this work, such as the approach for generating syntactic paraphrases as well as the approach for data collection of user preferences concerning syntactic forms via the audio channel. Similarly, the influence of syntactic forms in language perception and production was exploratorily investigated by means of driving simulation studies. In this context, potential parameters were identified as factors influencing the perception of syntactic forms in voice output and may serve as an entry point for adaptive concepts in SDS interaction as secondary task. Publications related to these experimental contributions are Stier and Sigloch (2019) and Stier *et al.* (2020b). Furthermore, syntactic complexity features of natural driver language were identified according to the respective complexity of the primary driving task. Related publications are Stier *et al.* (2020c,e). Subsequently, the developed adaptation strategy was evaluated in a real-life user study with respect

to user experience and driver distraction. Two pilot studies were additionally conducted to prove the reliability to estimate syntactic complexity and to demonstrate the feasibility to automatically detect contextual information in terms of personality and driving context from spoken language input. The publications related to these contributions are Stier *et al.* (2020a,d). Overall, the deduced adaptation strategy was shown to increase the user experience and thus represents a valuable improvement compared to a non-adaptive system (*cf.* Working Hypothesis 1), while Working Hypothesis 2 could not be confirmed in terms of a noticeable effect on a driver's cognitive load.

6.2 Suggestions for Future Work

The research work described in this thesis contributed to the identification parameters influencing the perception of syntactic forms in voice output and the characterization of spoken language in terms of syntactic complexity. Furthermore, the development of a syntactic adaptation strategy provided insights into the role of an individual user's personality traits and the respective driving situation concerning the user experience of an in-vehicle SDS. These findings allowed the realization of a first prototype and likewise offer the possibility for further research.

The goal of this research work was focused on an SDS user's personality and the driving condition. However, in the course of this research work, further user-side parameters have been exploratorily identified as factors influencing the perception of voice prompts in SDS interaction as secondary task while driving, such as the age, gender or linguistic experience. Unfortunately, a further reflection of these parameters was not possible within the scope of this thesis. Through the present results, however, it can be assumed that a more granular identification and expansion of the understanding of a user with respect to these factors would provide a valuable contribution to the goal of user-adaptive voice output. For instance, it is reasonable that the presented adaptation strategy could be adapted with the distinction of the age of extraverted users and cover the needs of the individual user more specifically. With regard to the syntactic design of voice output, in particular system-side parameters, such as the domain and question type, were further identified under the aspect of the interaction-defining situation. Also in this case, an extension of the understanding of the utilization context would be reasonable, for example by including further domains and dissolving the restriction to one-shot question-answer sequences that was assumed in this work. This could be achieved, for

example, by applying further vehicle-related domains, such as navigation or telephony, further dual-task scenarios, such as cooking according to instructions, or, in general, by considering multi-step dialogues.

The present work did not claim to examine voice output with respect to all linguistic levels. Instead, the focus was on the syntactic design and complexity of voice prompts. Future research is therefore necessary to also consider other aspects of human language in order to achieve the goal of natural interaction according to the human model. In particular, in the context of user and situation adaptivity, the inclusion of the semantic level with respect to the design of voice output could be investigated. It is conceivable that, for example, the use of different words influences the perception and user experience of voice output. For example, even the use of “voice assistant” instead of “dialog system” could represent a more intuitively understandable term in an explanation.

Although it was concluded from the present work that the perception of syntactic differences does not depend on driving complexity (based on the distinction between autonomous and manual driving), the conducted research nevertheless showed that the complexity of the driving situation has an influence on a driver’s speech production. With respect to future mobility, where an increasing degree of automation is assumed, it is expected that the adaptation strategy presented in this work is not limited to manual driving as a primary task with SAE level 0, but in general can also be applied in future in-vehicle voice interaction (*e.g.*, SAE level 5). By continuously evaluating the complexity of user language while driving by means of the relevant syntactic complexity components, voice output can consistently be adapted to the personal and situational requirements of an SDS user and road traffic. Nevertheless, this assumption requires verification in the context of further research, such as driving simulation studies within a more fine-grained differentiation of driving conditions.

In this thesis, an adaptation strategy for German-speaking SDS users was presented. However, it is probable that the user interaction and user experience will be different for other nationalities and cultures. It is therefore advisable to transfer the research conducted in this thesis to other cultural communities and to adapt the developed adaptation strategy accordingly in order to address culturally different needs in a more targeted way.

According to the goal of this work, the presented adaptation strategy was realized as a prototype and evaluated in the context of a real-life user study. Although the research steps and approaches carried out up to this point were validated and substantiated with the help of additional pilot studies, some weaknesses in the realization and implementation were revealed.

On the one hand, this concerns the functionality of the developed prototype. The computation of the syntactic complexity of user language is highly dependent on the accuracy of speech recognition. Although the employed ASR component was classified as sufficiently reliable for the context of this work, it can be assumed that for the purpose of recognizing spontaneous user speech, an improvement would be necessary in the long-term in order to ensure the most error-free possible application of the adaptation strategy. Furthermore, the prototype is based on currently naive approaches to compute syntactic complexity and identify contextual information. With the help of current research advances, these aspects could be improved, for example, through the employment of machine learning. A second weakness concerns the implemented user study for evaluation. Here, the small number of subjects in general and their number per user cluster have to be mentioned, the latter involving a certain difficulty in the sense of a “black box,” as in an anonymous subject acquisition a balanced distribution of the required user groups is hardly achievable. Future research is needed to specifically address these aspects. Nonetheless, the user- and situation-adaptive strategy for the syntactic design of voice output that was presented in this thesis with a focus on user experience from the user’s perspective is considered to represent a further building block for the long-term goal of intuitive, natural, and conversational SDS interaction in dual-task scenarios according to the human model.

Appendix A

Materials of the Studies on Language Perception

This appendix chapter provides the materials employed for the studies on language perception. It furthermore includes additional information concerning the performed analyses.

A.1 Preparation of Syntactic Paraphrases

This appendix section contains manually created syntactic paraphrases for the two vehicle-related domains DAS and COP. While the DAS paraphrases describe the functionality of driving assistants as answer to the question “What is...?,” the COP paraphrases provide descriptions of individual comfort functions. The paraphrases were employed as explanatory voice prompts in the following investigations and constituted the basis for the studies on language perception.

A.2 Pilot Studies

This appendix section provides the material of two pilot studies, which served as preliminary investigations for further research. The first one aimed at validating the manual preparation approach of syntactic paraphrases (Appendix A.2.1), whereas the second one investigated whether SDS users are generally aware of syntactic forms in voice output (Appendix A.2.2).

Table A.1: Syntactic paraphrases generated for the domain DAS.

DAS function	MCV	FCV	NCV	RCV
Abstands-Assistent (*Space Assist*)	Der aktive Abstands-Assistent kann Sie bei zu dichtem Auffahren warnen. Er bremst Ihr Fahrzeug gegebenenfalls ab, um den Abstand zu vorausfahrenden Fahrzeugen zu regeln.	Der aktive Abstands-Assistent kann Sie bei zu dichtem Auffahren warnen und bremst Ihr Fahrzeug gegebenenfalls ab, um den Abstand zu vorausfahrenden Fahrzeugen zu regeln.	Der aktive Abstands-Assistent kann Sie bei zu dichtem Auffahren warnen und regelt durch Abbremsen Ihres Fahrzeugs den Abstand zu vorausfahrenden Fahrzeugen.	Der aktive Abstands-Assistent, der Sie bei zu dichtem Auffahren warnen kann, bremst Ihr Fahrzeug gegebenenfalls ab, um den Abstand zu vorausfahrenden Fahrzeugen zu regeln.
Nothalt-Assistent (*Emergency Stop Assist*)	Der aktive Nothalt-Assistent kann Sie bei dauerhafter Ablenkung warnen. Er bremst Ihr Fahrzeug kontrolliert bis zum Stillstand ab, um eine Kollision zu verhindern.	Der aktive Nothalt-Assistent kann Sie bei dauerhafter Ablenkung warnen und bremst Ihr Fahrzeug kontrolliert bis zum Stillstand ab, um eine Kollision zu verhindern.	Der aktive Nothalt-Assistent kann Sie bei dauerhafter Ablenkung warnen und verhindert durch Abbremsen Ihres Fahrzeugs eine Kollision.	Der aktive Nothalt-Assistent, der Sie bei dauerhafter Ablenkung warnen kann, bremst Ihr Fahrzeug kontrolliert bis zum Stillstand ab, um eine Kollision zu verhindern.
Spurhalte-Assistent (*Lane Keeping Assist*)	Der aktive Spurhalte-Assistent kann Sie bei unbeabsichtigtem Verlassen der Fahrspur warnen. Er bremst eigenständig, um Ihr Fahrzeug zurück in die Spur zu führen.	Der aktive Spurhalte-Assistent kann Sie bei unbeabsichtigtem Verlassen der Fahrspur warnen und bremst eigenständig, um Ihr Fahrzeug zurück in die Spur zu führen.	Der aktive Spurhalte-Assistent kann Sie bei unbeabsichtigtem Verlassen der Fahrspur warnen und führt Ihr Fahrzeug durch eigenständiges Bremsen zurück in die Spur.	Der Aktive Spurhalte-Assistent, der Sie bei unbeabsichtigtem Verlassen der Fahrspur warnen kann, bremst eigenständig, um Ihr Fahrzeug zurück in die Spur zu führen.
Totwinkel-Assistent (*Blind Spot Assist*)	Der aktive Totwinkel-Assistent kann Sie bei einem Spurwechsel vor Fahrzeugen im toten Winkel warnen. Er bremst Ihr Fahrzeug eigenständig ab, um eine Kollision zu vermeiden.	Der aktive Totwinkel-Assistent warnt Sie bei einem Spurwechsel vor Fahrzeugen im toten Winkel und bremst Ihr Fahrzeug eigenständig ab, um eine Kollision zu vermeiden.	Der aktive Totwinkel-Assistent kann Sie bei einem Spurwechsel vor Fahrzeugen im toten Winkel warnen und vermeidet durch eigenständiges Abbremsen Ihres Fahrzeugs eine Kollision.	Der aktive Totwinkel-Assistent, der Sie bei einem Spurwechsel vor Fahrzeugen im toten Winkel warnen kann, bremst Ihr Fahrzeug eigenständig ab, um eine Kollision zu vermeiden.

Table A.2: Syntactic paraphrases generated for the domain COP.

COP function	MCV	FCV	NCV	RCV
Behaglichkeit (*Well-being*)	Das Programm Behaglichkeit kann Ihrem Komfort in angespannten Fahrsituationen dienen. Es erzeugt durch eine wärmende Massage ein entspanntes Spa-Feeling, um Ihr Wolbefinden zu steigern.	Das Programm Behaglichkeit kann Ihrem Komfort in angespannten Fahrsituationen dienen und erzeugt durch eine wärmende Massage ein entspanntes Spa-Feeling, um Ihr Wohlbefinden zu steigern.	Das Programm Behaglichkeit kann Ihrem Komfort in angespannten Fahrsituationen dienen und steigert ihr Wohlbefinden durch eine wärmende Massage und das Erzeugen eines enstpannten Spa-Feelings.	Das Programm Behaglichkeit, das Ihrem Komfort in angespannten Fahrsituationen dienen kann, erzeugt durch eine wärmende Massage ein entspanntes Spa-Feeling, um Ihr Wolbefinden zu steigern.
Freude (*Joy*)	Das Programm Freude kann Ihrem Komfort in ermüdenden Fahrsituationen dienen. Es nutzt eine aktivierende Sitzmassage und Musik, um eine positive Stimmung zu erzeugen.	Das Programm Freude kann Ihrem Komfort in ermüdenden Fahrsituationen dienen und nutzt eine aktivierende Sitzmassage und Musik, um eine positive Stimmung zu erzeugen.	Das Programm Freude kann Ihrem Komfort in ermüdenden Fahrsituationen dienen und erzeugt eine positive Stimmung durch die Nutzung einer aktivierenden Sitzmassage und Musik.	Das Programm Freude, das Ihrem Komfort in ermüdenden Fahrsituationen dienen kann, nutzt eine aktivierende Sitzmassage und Musik, um eine positive Stimmung zu erzeugen.
Vitalität (*Vitality*)	Das Programm Vitalität kann Ihrem Komfort in monotonen Fahrsituationen dienen. Es nutzt anregendes Licht und Musik, um eine belebende Wirkung zu erzeugen.	Das Programm Vitalität kann Ihrem Komfort in monotonen Fahrsituationen dienen und nutzt anregendes Licht und Musik, um eine belebende Wirkung zu erzeugen.	Das Programm Vitalität kann Ihrem Komfort in monotonen Fahrsituationen dienen und erzeugt eine belebende Wirkung durch die Nutzung von anregendem Licht und Musik.	Das Programm Vitalität, das Ihrem Komfort in monotonen Fahrsituationen dienen kann, nutzt anregendes Licht und Musik, um eine belebende Wirkung zu erzeugen.
Wärme (*Warmth*)	Das Programm Wärme kann Ihrem Komfort in belastenden Fahrsituationen dienen. Es erzeugt durch beheizte Sitze eine wohlige Wärme, um für ein gemütliches Ambiente zu sorgen.	Das Programm Wärme kann Ihrem Komfort in belastenden Fahrsituationen dienen und erzeugt durch beheizte Sitze wohlige Wärme, um für ein gemütliches Ambiente zu sorgen.	Das Programm Wärme kann Ihrem Komfort in belastenden Fahrsituationen dienen und sorgt für ein gemütliches Ambiente durch die Beheizung der Sitze und das Erzeugen wohliger Wärme.	Das Programm Wärme, das Ihrem Komfort in belastenden Fahrsituationen dienen kann, erzeugt durch beheizte Sitze eine wohlige Wärme, um für ein gemütliches Ambiente zu sorgen.

A.2.1 Validation of the Preparation Approach

This questionnaire is reprinted with kind permission of my co-author Ellen Sigloch.

Antworten zu verschiedenen Fahrassistenten.	stimme nicht zu	neutral	stimme zu
Antwort 1: Die Antwort empfinde ich als passend auf meine Frage.			
Antwort 2: Die Antwort empfinde ich als passend auf meine Frage.			
Antwort 3: Die Antwort empfinde ich als passend auf meine Frage.			
Antwort 4: Die Antwort empfinde ich als passend auf meine Frage.			

Antwort 1 – Antwort 2 – Antwort 3 – Antwort 4	stimme nicht zu	neutral	stimme zu
Ich erkenne eine wiederkehrende Struktur.			
Die Inhalte scheinen mir einheitlich formuliert.			

Antwort 1	stimme nicht zu	neutral	stimme zu
Die Ausdrucksweise empfinde ich als natürlichsprachlich.			
Der Inhalt ist leicht verständlich.			
Antwort 2	stimme nicht zu	neutral	stimme zu
Die Ausdrucksweise empfinde ich als natürlichsprachlich.			
Der Inhalt ist leicht verständlich.			
Antwort 3	stimme nicht zu	neutral	stimme zu
Die Ausdrucksweise empfinde ich als natürlichsprachlich.			
Der Inhalt ist leicht verständlich.			
Antwort 4	stimme nicht zu	neutral	stimme zu
Die Ausdrucksweise empfinde ich als natürlichsprachlich.			
Der Inhalt ist leicht verständlich.			

A.2.2 Investigating the Level of Consciousness

Versuchspersonennummer (siehe Terminplan Probandenmanagement)

VP

[Bitte auswählen]

Audio-Erklärung zur Beurteilung vorspielen

VL: "Ich würde Ihnen gerne hintereinander zwei Dateien vorspielen. Sie hören jeweils eine Erklärungen zu einem Fahrassistenzsystem. Bitte lassen Sie sich nicht von der Stimme oder eventuellen Aussprachefehlern beeinflussen.

Achten Sie hier bitte nicht auf den Inhalt, sondern vielmehr auf die Formulierung und die Art, *wie* etwas erklärt wird."

Ist Ihnen an der Formulierung der Erklärung etwas Besonderes aufgefallen?
(Satzstellung / Grammatik)

- ○ Nein
- ○ Haupt-/Relativsatz
- ○ Sonstiges
- ○ unklar/ nicht eindeutig

Audio-Erklärung zur Beurteilung vorspielen (nur 1x vorspielen!)

Ist Ihnen an der Formulierung der Erklärung etwas Besonderes aufgefallen?
(Satzstellung / Grammatik)

- ○ Nein
- ○ Haupt-/Relativsatz
- ○ Sonstiges
- ○ unklar/ nicht eindeutig

Sind Ihnen Unterschiede zwischen den Erklärungen aufgefallen?

- ○ Nein
- ○ Haupt-/Relativsatz
- ○ Sonstiges
- ○ unklar/ nicht eindeutig

Welche Erklärung gefällt Ihnen persönlich besser?

- ○ Erklärung 1 (Audio)
- ○ Erklärung 2 (Audio)
- ○ beide gleich

Warum?
OPTIONAL

- Empfinden Sie die Formulierung als natürlicher?
- Glauben Sie die Formulierung ist so leichter zu verstehen?

Hinweis: Angabe der Präferenz zum aktuellen Zeitpunkt, d. h. Proband soll sich nicht vorstellen, was für die Fahrt besser wäre, sondern was er in diesem Moment als besser/schöner/natürlicher/intuitiver empfindet!)

Unterschied 1 – 2

Text-Erklärung zur Beurteilung vorlegen (einzeln vorlegen und wieder wegnehmen!)

VL: "Jetzt möchte ich Ihnen gerne hintereinander zwei kurze Texte zeigen. Auch hier geht es jeweils um die Erklärung eines Fahrerassistenzsystems.

Achten Sie hier bitte nicht auf den Inhalt, sondern vielmehr auf die Formulierung und die Art, *wie* etwas erklärt wird."

Ist Ihnen an der Formulierung der Erklärung etwas Besonderes aufgefallen?
(Satzstellung / Grammatik)

- ○ Nein
- ○ Haupt-/Relativsatz
- ○ Sonstiges
- ○ unklar/ nicht eindeutig

Text-Erklärung zur Beurteilung vorlegen (einzeln vorlegen und wieder wegnehmen!)

Ist Ihnen an der Formulierung der Erklärung etwas Besonderes aufgefallen?
(Satzstellung / Grammatik)

- ○ Nein
- ○ Haupt-/Relativsatz
- ○ Sonstiges
- ○ unklar/ nicht eindeutig

Sind Ihnen Unterschiede zwischen den Erklärungen aufgefallen?

- ○ Nein
- ○ Haupt-/Relativsatz
- ○ Sonstiges
- ○ unklar/ nicht eindeutig

Welche Erklärung gefällt Ihnen persönlich besser?

- ○ Erklärung 3 (Text)
- ○ Erklärung 4 (Text)
- ○ beide gleich

Erklärung der Unterschiede

Relativsatz - zwei Hauptsätze

Welche Erklärung gefällt Ihnen persönlich besser?

- ○ Erklärung 5 (Audio)
- ○ Erklärung 6 (Audio)
- ○ beide gleich

Warum?
OPTIONAL

- Empfinden Sie die Formulierung als natürlicher?

- Glauben Sie die Formulierung ist so leichter zu verstehen?

Hinweis: Angabe der Präferenz zum aktuellen Zeitpunkt, d. h. Proband soll sich nicht vorstellen, was für die Fahrt besser wäre, sondern was er in diesem Moment als besser/schöner/natürlicher/intuitiver empfindet!)

Unterschied 5 – 6 (Audio)	

A.2.3 Syntactic Paraphrases for Presentation in Audio Form

This appendix section contains the syntactic paraphrases that were synthesized and subsequently presented to the participants of the second pilot study in order to examine whether they identified the syntactic differences via audio. Each participant randomly received one of the four paraphrase pairs. Thereby, the order of syntactic forms was randomized for two different DAS functions.

	Audio 1		Audio 2	
Set 1	**Was ist der Abstands-Assistent?**	Der Aktive Abstands-Assistent kann einen sicheren Abstand zum vorausfahrenden Fahrzeug halten. So verringert er das Risiko von Auffahrunfällen.	**Was ist der Brems-Assistent?**	Der Aktive Brems-Assistent, der mithilfe der Abstandswarnfunktion das Risiko einer Kollision erkennen kann, vermeidet so die Gefahr von Auffahrunfällen.
Set 2	**Was ist der Brems-Assistent?**	Der Aktive Brems-Assistent kann mithilfe der Abstandswarnfunktion das Risiko einer Kollision erkennen. So vermeidet er die Gefahr von Auffahrunfällen.	**Was ist der Spurhalte-Assistent?**	Der Aktive Spurhalte-Assistent, der Sie vor unbeabsichtigten Spurwechseln schützen kann, verringert so die Gefahr einer seitlichen Kollision.
Set 3	**Was ist der Totwinkel-Assistent?**	Der Aktive Totwinkel-Assistent, der Fahrzeuge im toten Winkel erkennen kann, vermeidet so das Risiko von Kollisionen mit anderen Fahrzeugen.	**Was ist der Spurhalte-Assistent?**	Der Aktive Spurhalte-Assistent kann Sie vor unbeabsichtigten Spurwechseln warnen und schützen. So verringert er die Gefahr einer seitlichen Kollision.
Set 4	**Was ist der Abstands-Assistent?**	Der Aktive Abstands-Assistent, der einen sicheren Abstand zum vorausfahrenden Fahrzeug halten kann, verringert so das Risiko von Auffahrunfällen.	**Was ist der Totwinkel-Assistent?**	Der Aktive Totwinkel-Assistent kann Fahrzeuge im toten Winkel erkennen. So vermeidet er das Risiko von Kollisionen mit anderen Fahrzeugen.

A.2.4 Syntactic Paraphrases for Presentation in Text Form

In this section, the text samples are provided that were presented to the participants of the second pilot study to examine whether they identified the syntactic differences between the respective two paraphrases based on their textual representation. They were additionally used to explain the syntactic differences between the paratactic alignment of sentences in one variant and a nested, subject-oriented relative clause in the other variant. According to the set chosen in Appendix A.2.3, each participant was presented with the corresponding paraphrase pair in the same sequence of syntactic forms. Thus, each participant was presented with four different DAS functions to direct the focus of attention to the syntactic differences and away from the content level.

Set 1

Text 1		Text 2	
Was ist der Spurhalte-Assistent?	Der Aktive Spurhalte-Assistent kann Sie vor unbeabsichtigten Spurwechseln warnen und schützen. So verringert er die Gefahr einer seitlichen Kollision.	**Was ist der Totwinkel-Assistent?**	Der Aktive Totwinkel-Assistent, der Fahrzeuge im toten Winkel erkennen kann, vermeidet so das Risiko von Kollisionen mit anderen Fahrzeugen.

Set 2

Text 1		Text 2	
Was ist der Totwinkel-Assistent?	Der Aktive Totwinkel-Assistent kann Fahrzeuge im toten Winkel erkennen. So vermeidet er das Risiko von Kollisionen mit anderen Fahrzeugen.	**Was ist der Abstands-Assistent?**	Der Aktive Abstands-Assistent, der einen sicheren Abstand zum vorausfahrenden Fahrzeug halten kann, verringert so das Risiko von Auffahrunfällen.

Set 3

Text 1		Text 2	
Was ist der Brems-Assistent?	Der Aktive Brems-Assistent, der mithilfe der Abstandswarnfunktion das Risiko einer Kollision erkennen kann, vermeidet so die Gefahr von Auffahrunfällen.	**Was ist der Abstands-Assistent?**	Der Aktive Abstands-Assistent kann einen sicheren Abstand zum vorausfahrenden Fahrzeug halten. So verringert er das Risiko von Auffahrunfällen.

Set 4

Text 1		Text 2	
Was ist der Spurhalte-Assistent?	Der Aktive Spurhalte-Assistent, der Sie vor unbeabsichtigten Spurwechseln schützen kann, verringert so die Gefahr einer seitlichen Kollision.	**Was ist der Brems-Assistent?**	Der Aktive Brems-Assistent kann mithilfe der Abstandswarnfunktion das Risiko einer Kollision erkennen. So vermeidet er die Gefahr von Auffahrunfällen.

A.3 First User Study on the Influence of Syntax

This appendix section contains the materials of the first user study on the influence of syntactic forms in in-vehicle voice output. First, the preliminary questionnaire (Appendix A.3.1) and a description of the study content for the introduction of the study participants (Appendix A.3.2) are presented, followed by the evaluation scale (Appendix A.3.3). Appendix A.3.4 then provides an overview of the dialog tasks presented to the participants during the driving simulation study. Finally, the results of the evaluation are provided (Appendix A.3.5).

A.3.1 Pre-Survey

The purpose of this questionnaire was to measure individual user characteristics. Besides general demographic and experience-related questions, this survey includes the questionnaire by Karrer *et al.* (2009) concerning a participant's technical affinity (section 3) as well as the questionnaire by Rammstedt and Danner (2016) to measure Big Five Personality traits (section 4).

Vorbefragung

1. Demographische Daten

1a. Alter _____

1b. Geschlecht

□ männlich □ weiblich

1c. Höchster Bildungsabschluss

□ Volksschul-/Hauptschulabschluss
□ Mittlere Reife
□ Abitur/Fachabitur
□ Abgeschlossene Berufsausbildung
□ Hochschul-/Fachhochschulabschluss
□ Promotion
□ anderer: _____

1d. Aktuelle Tätigkeit

□ nicht arbeitstätig
□ Angestellt
□ Selbständig
□ in Ausbildung (Schule, Studium, etc.)
□ Hausfrau/Hausmann
□ Rentner/ Pensionär
□ andere Tätigkeit

2. Nutzung und Kenntnisse in der Bedienung von Sprachassistenten

2a. Wie bewerten Sie selbst Ihre linguistischen Kenntnisse?

sehr gering	□1	□2	□3	□4	□5	***sehr hoch***

2b. Wie bewerten Sie selbst Ihre Vorerfahrung mit Fahrassistenzsystemen?

sehr gering	□1	□2	□3	□4	□5	***sehr hoch***

2c. Wie bewerten Sie selbst Ihre Vorerfahrung mit ENERGIZING Comfort?

sehr gering	□1	□2	□3	□4	□5	***sehr hoch***

2d. Wie häufig nutzen Sie eine Sprachsteuerung?

nie	□1	□2	□3	□4	□5	***sehr häufig***	□0	***Ich habe keine.***

2e. Wie zufrieden sind Sie mit Ihrer Sprachbedienung?

gar nicht zufrieden	□1	□2	□3	□4	□5	***sehr zufrieden***	□0	***nicht beurteilbar***

3. TECHNIKBEZOGENE SELBSTEINSCHÄTZUNG

Inwieweit treffen die folgenden Aussagen auf Sie persönlich zu? Lesen Sie bitte jede Aussage aufmerksam durch und entscheiden Sie dann, wie sehr Sie der jeweiligen Aussage zustimmen. Bitte beantworten Sie jede Aussage spontan und wahrheitsgemäß.

Unter dem Begriff „elektronische Geräte" verstehen wir hier Geräte, wie Computer, Handys, Digitalkameras, Geldautomaten, sowie neue Systeme im Auto wie Navigationssysteme. Nicht gemeint sind Werkzeuge (Bohrmaschine, Hammer), Haushaltsgeräte (Toaster, Wasserkocher), Fahrzeuge oder Fahrzeugmotoren.

		sehr unzutreffend	***eher unzutreffend***	***teils/ teils***	***eher zutreffend***	***sehr zutreffend***
1.	Ich liebe es, neue elektronische Geräte zu besitzen.	□	□	□	□	□
2.	Elektronische Geräte machen krank.	□	□	□	□	□
3.	Ich gehe gern in den Fachhandel für elektronische Geräte.	□	□	□	□	□
4.	Ich habe bzw. hätte Verständnisprobleme beim Lesen von Elektronik- und Computerzeitschriften.	□	□	□	□	□
5.	Elektronische Geräte ermöglichen einen hohen Lebensstandard.	□	□	□	□	□
6.	Elektronische Geräte führen zu geistiger Verarmung.	□	□	□	□	□
7.	Elektronische Geräte machen vieles umständlicher.	□	□	□	□	□
8.	Ich informiere mich über elektronische Geräte, auch wenn ich keine Kaufabsicht habe.	□	□	□	□	□
9.	Elektronische Geräte machen unabhängig.	□	□	□	□	□
10.	Es macht mir Spaß, ein elektronisches Gerät auszuprobieren.	□	□	□	□	□
11.	Elektronische Geräte erleichtern mir den Alltag.	□	□	□	□	□
12.	Elektronische Geräte erhöhen die Sicherheit.	□	□	□	□	□
13.	Elektronische Geräte verringern den persönlichen Kontakt zwischen den Menschen.	□	□	□	□	□
14.	Ich kenne die meisten Funktionen der elektronischen Geräte, die ich besitze.	□	□	□	□	□
15.	Ich bin begeistert, wenn ein neues elektronisches Gerät auf den Markt kommt.	□	□	□	□	□
16.	Elektronische Geräte verursachen Stress.	□	□	□	□	□
17.	Ich kenne mich im Bereich elektronischer Geräte aus.	□	□	□	□	□
18.	Es fällt mir leicht, die Bedienung eines elektronischen Geräts zu lernen.	□	□	□	□	□
19.	Elektronische Geräte helfen, an Informationen zu gelangen.	□	□	□	□	□

4. PERSÖNLICHKEITSBEZOGENE SELBSTEINSCHÄTZUNG

Inwieweit treffen die folgenden Aussagen auf Sie persönlich zu? Lesen Sie bitte jede Aussage aufmerksam durch und entscheiden Sie dann, wie sehr Sie der jeweiligen Aussage zustimmen. Bitte beantworten Sie jede Aussage spontan und wahrheitsgemäß.

	Ich …	***sehr unzutreffend***	***eher unzutreffend***	***teils/ teils***	***eher zutreffend***	***sehr zutreffend***
1.	… bin gesprächig, unterhalte mich gern.	□	□	□	□	□
2.	… neige dazu, andere zu kritisieren.	□	□	□	□	□
3.	… erledige Aufgaben gründlich.	□	□	□	□	□
4.	… bin deprimiert, niedergeschlagen.	□	□	□	□	□
5.	… bin originelle, entwickle neue Ideen.	□	□	□	□	□
6.	… bin eher zurückhaltend, reserviert.	□	□	□	□	□
7.	… bin hilfsbereit und selbstlos gegenüber anderen.	□	□	□	□	□
8.	… bin manchmal unsorgfältig und schluderig.	□	□	□	□	□
9.	… bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	□	□	□	□	□
10.	… bin vielseitig interessiert.	□	□	□	□	□
11.	… bin voller Energie und Tatendrang.	□	□	□	□	□
12.	… bin häufig in Streitereien verwickelt.	□	□	□	□	□
13.	… arbeite zuverlässig und gewissenhaft.	□	□	□	□	□
14.	… reagiere leicht angespannt.	□	□	□	□	□
15.	… bin tiefsinnig, denke gerne über Sachen nach.	□	□	□	□	□
16.	… bin begeisterungsfähig und kann andere leicht mitreißen.	□	□	□	□	□
17.	… bin nicht nachtragend, vergebe anderen leicht.	□	□	□	□	□
18.	… bin eher unordentlich.	□	□	□	□	□
19.	… mache mir viele Sorgen.	□	□	□	□	□
20.	… habe eine aktive Vorstellungskraft, bin fantasievoll.	□	□	□	□	□
21.	… bin eher der „stille Typ“, wortkarg.	□	□	□	□	□
22.	… schenke anderen Vertrauen, glaube an das Gute im Menschen.	□	□	□	□	□
23.	… bin bequem, neige zur Faulheit.	□	□	□	□	□

	Ich …	***sehr unzutreffend***	***eher unzutreffend***	***teils/ teils***	***eher zutreffend***	***sehr zutreffend***
24.	… bin emotional ausgeglichen, nicht leicht aus der Fassung zu bringen.	□	□	□	□	□
25.	… bin erfinderisch und einfallsreich.	□	□	□	□	□
26.	… bin durchsetzungsfähig, energisch.	□	□	□	□	□
27.	… kann mich kalt und distanziert verhalten.	□	□	□	□	□
28.	… harre aus (und arbeite weiter), bis die Aufgabe fertig ist.	□	□	□	□	□
29.	… kann launisch sein, habe schwankende Stimmungen.	□	□	□	□	□
30.	… schätze künstlerische und ästhetische Eindrücke.	□	□	□	□	□
31.	… bin manchmal schüchtern und gehemmt.	□	□	□	□	□
32.	… bin rücksichtsvoll zu anderen, einfühlsam.	□	□	□	□	□
33.	… bin tüchtig und arbeite flott.	□	□	□	□	□
34.	… bleibe ruhig, selbst in Stresssituationen.	□	□	□	□	□
35.	… mag es, wenn Aufgaben routinemäßig zu erledigen sind.	□	□	□	□	□
36.	… gehe aus mir heraus, bin gesellig.	□	□	□	□	□
37.	… kann mich schroff und abweisend anderen gegenüber verhalten.	□	□	□	□	□
38.	… mache Pläne und führe sie auch durch.	□	□	□	□	□
39.	… werde leicht nervös und unsicher.	□	□	□	□	□
40.	… stelle gerne Überlegungen an, spiele mit abstrakten Ideen.	□	□	□	□	□
41.	… habe nur wenig künstlerisches Interesse.	□	□	□	□	□
42.	… verhalte mich kooperativ, ziehe Zusammenarbeit dem Wettbewerb vor.	□	□	□	□	□
43.	… bin leicht ablenkbar, bleibe nicht bei der Sache.	□	□	□	□	□
44.	… kenne mich gut in Musik, Kunst oder Literatur aus.	□	□	□	□	□
45.	… habe oft Krach mit anderen.	□	□	□	□	□

A.3.2 Introduction of the Study Content

INHALT DER STUDIE

Das Ziel dieser Studie ist festzustellen, ob und inwiefern der individuelle Gebrauch von Sprache in der Gestaltung von Sprachdialogausgaben zu berücksichtigen ist. Hierfür werden Sie gebeten, unter verschiedenen Fahrbedingungen mit einem Sprachassistenten als Fahrer in Interaktion zu treten. Die Fahrbedingung wird durch den Einsatz eines autonomen Fahrmodus variiert, d.h. einen Teil der Fahrt werden Sie selbst die Steuerung des Fahrzeugs übernehmen und einen Teil der Fahrt wird das Fahrzeug autonom (selbständig) fahren.

Ihre Fahrt wird bei Tag auf einer Autobahn stattfinden, wobei Sie dazu aufgefordert sind eine möglichst konstante Geschwindigkeit von 100km/h einzuhalten.

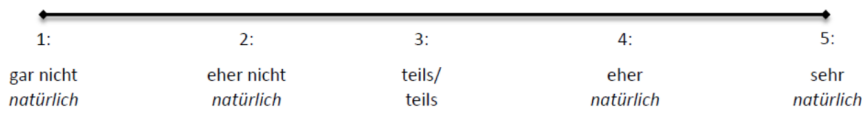
Während der Fahrt werden Ihnen verschiedene Aufgaben aus den Bereichen *Fahrassistenzsysteme* und *Energizing Comfort Programme* präsentiert, auf deren Basis Sie dem Sprachassistenten eine Frage stellen sollen. Bevor Sie dem Sprachassistenten eine Frage stellen können, aktivieren Sie ihn bitte durch „Hallo Mercedes." Sie werden anschließend darum gebeten, die vom Sprachassistenten gelieferte Antwort hinsichtlich ihrer *Natürlichkeit* zu bewerten.

Bitte hören Sie den vom Sprachassistenten gelieferten Antworten aufmerksam zu, denn Sie werden zu verschiedenen Zeitpunkten der Fahrt angerufen und gebeten, den zuletzt gehörten Inhalt wiederzugeben.

A.3.3 Evaluation Scale

This appendix section presents the evaluation scale that was used in this user study. Besides the scale, the participants were introduced to concrete criteria they should consider in their evaluation of voice prompts. This particularly included the aspects of the perceived naturalness and comprehensibility of voice output. The participants were asked to evaluate these aspects jointly on a 5-point Likert scale.

Auf einer Skala von 1 (gar nicht *natürlich*) bis 5 (sehr *natürlich*), wie bewerten Sie die gehörte Antwort?



Kriterien:

- **Natürlichkeit** Würden Sie die Antwort auch so formulieren?
- **Intuitivität** Ist die Antwort intuitiv und sofort verständlich?
- **Verständlichkeit** Ist die Antwort in ihrer Formulierung gut verständlich?

A.3.4 Dialog Tasks

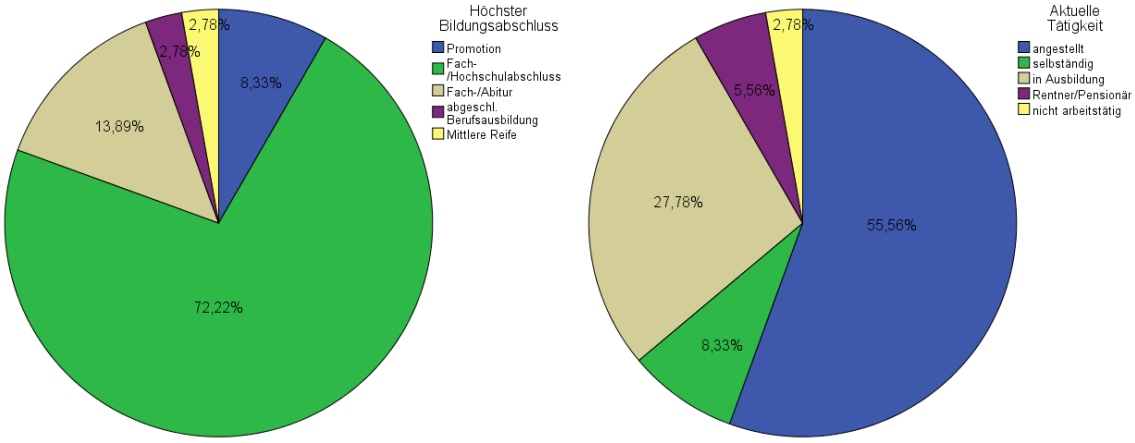
This appendix section contains the dialog tasks used in the user study. During the driving simulation, they were projected onto the head-unit to indicate the question a participants should formulate. Based on the *Brems-Assistent* (“Brake Assist”), the participants were explained that each dialog task consisted of a picture and name representing the vehicle function to be inquired and the question type What.

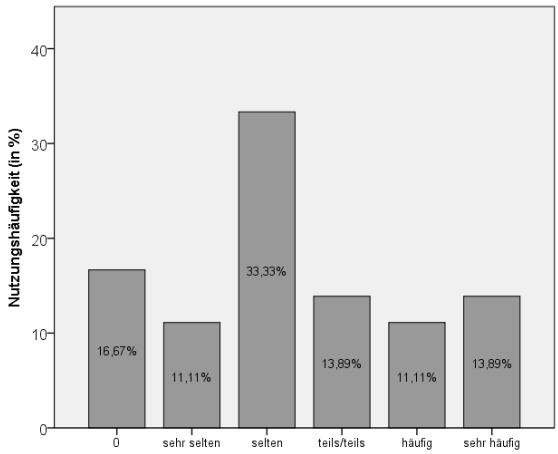
The graphics have been removed due to copyright limitations.

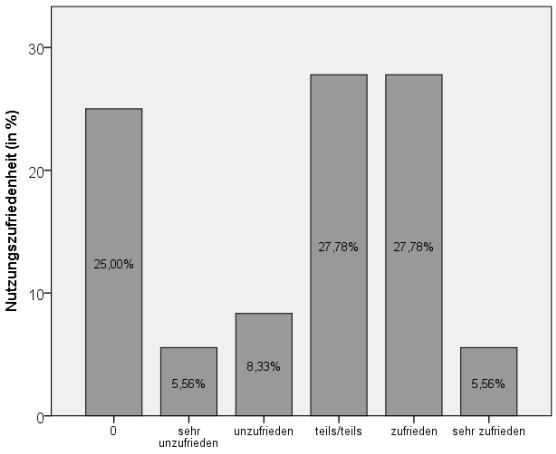
A.3.5 Results of the first User Study on the Influence of Syntax

This appendix section provides an overview of the results obtained for the first user study on the influence of syntactic forms in voice output.

The following figures contain supplementary results about the participants from the pre-survey. The majority of participants reported having a university degree and being in a salaried or apprenticeship position at the time of the study. Overall, the participants indicated that they generally used spoken dialog systems rather infrequently, but with an average level of satisfaction.







A.3.5.1 Generalized Linear Mixed Model

In the following, the results of the explorative generalized linear mixed model are presented.

Modellübersicht

Ziel		naturalness
Messniveau		Ordinal
Wahrscheinlichkeitsverteilung		Multinomial
Verknüpfungsfunktion		Logit (kumulativ)
Informationskriterium	Akaike (korrigiert)	15450,904
	Bayes	15455,156

Informationskriterien beruhen auf der -2 Log-Likelihood (15448,896) und dienen zum Modellvergleich. Modelle mit kleineren Werten für Informationskriterien weisen eine bessere Anpassung auf.

Zusammenfassung der Fallverarbeitung

	N	Prozent
Eingeschlossen	576	100,00%
Ausgeschlossen	0	0,00%
Gesamtergebnis	576	100,00%

Klassifikation
Gesamtprozent korrekt = 63,9%[a]

		Vorhergesagt				
Beobachtet		sehr unnatürlich	eher unnatürlich	teils/teils	eher natürlich	sehr natürlich
sehr unnatürlich	Anzahl	0	1	1	4	0
	% in 'Beobachtet'	0,00%	16,70%	16,70%	66,70%	0,00%
eher unnatürlich	Anzahl	0	0	26	11	0
	% in 'Beobachtet'	0,00%	0,00%	70,30%	29,70%	0,00%
teils/teils	Anzahl	0	2	70	52	0
	% in 'Beobachtet'	0,00%	1,60%	56,50%	41,90%	0,00%
eher natürlich	Anzahl	0	0	33	190	28
	% in 'Beobachtet'	0,00%	0,00%	13,10%	75,70%	11,20%
sehr natürlich	Anzahl	0	0	2	48	108
	% in 'Beobachtet'	0,00%	0,00%	1,30%	30,40%	68,40%

a. Ziel: naturalness

Kovarianzparameter-Übersicht

Kovarianzparameter		
Kovarianzparameter	Residualeffekt	0
	Zufällige Effekte	1
Design-Matrix-Spalten	Feste Effekte	85
	Zufällige Effekte	1[a]
Gemeinsame Subjekte		36

Gemeinsame Subjekte beruhen auf den Subjektspezifikationen für den Residualeffekt und die zufälligen Effekte und dienen dazu, die Daten aufzuteilen, um eine bessere Leistungsfähigkeit zu erreichen.

a. Dies ist die Anzahl an Spalten pro gemeinsamem Subjekt.

Block für zufällige Effekte 1

Block für zufällige Effekte	Konstanter Term
Konstanter Term	3,036

Kovarianzstruktur: Skalierte Identität

Subjektspezifikation: ID

Zufälliger Effekt

					95% Konfidenzintervall	
Zufälliger Effekt Kovarianz	Schätzer	Standard Fehler	Z	Sig.	Unterer Wert	Oberer Wert
Varianz	3,036	1,333	2,278	0,023	1,284	7,178

Kovarianzstruktur: Skalierte Identität

Subjektspezifikation: ID

Feste Effekte[a]

Quelle	F	df1	df2	Sig.
Korrigiertes Modell	3,851E+12	36	523	0,000
complexity	0,104	1	523	0,747
domain	21,686	1	523	0,000
sentence_type	8,962	1	523	0,003
complexity * sentence_type	0,098	1	523	0,754
domain * sentence_type	1,429	1	523	0,233
age_groups	5,444	3	523	0,001
gender	0,002	1	523	0,966
experience_linguistics_kat	4,012	2	523	0,019
experience_fas_kat	4,038	2	523	0,018
experience_enc_kat	17,436	2	523	0,000
bfi_openness_kat	2,595	1	523	0,108
bfi_conscientiousness_kat	8,853	1	523	0,003
bfi_extraversion_kat	3,312	1	523	0,069
bfi_agreeableness_kat	4,382	1	523	0,037
bfi_neuroticism_kat	32,862	2	523	0,000
ta_competence_kat	3,38	1	523	0,067
ta_enthusiasm_kat	7,632	2	523	0,001
ta_positive_attitude_kat	1,586	2	523	0,206
ta_negative_attitude_kat	0,28	1	523	0,597
age_groups * sentence_type	6,21	3	523	0,000
gender * sentence_type	0,857	1	523	0,355
experience_linguistics_kat * sentence_type	1,114	2	523	0,329
experience_fas_kat * sentence_type	7,057	2	523	0,001
experience_enc_kat * sentence_type	2,211	2	523	0,111
bfi_openness_kat * sentence_type	0,448	1	523	0,504
bfi_conscientiousness_kat * sentence_type	0,102	1	523	0,749
bfi_extraversion_kat * sentence_type	5,143	1	523	0,024
bfi_agreeableness_kat * sentence_type	2,899	1	523	0,089
bfi_neuroticism_kat * sentence_type	9,421	2	523	0,000
ta_competence_kat * sentence_type	12,739	1	523	0,000
ta_enthusiasm_kat * sentence_type	0,398	2	523	0,672
ta_positive_attitude_kat * sentence_type	2,827	2	523	0,060
ta_negative_attitude_kat * sentence_type	7,756	1	523	0,006

Wahrscheinlichkeitsverteilung: Multinomial

Verknüpfungsfunktion: Logit (kumulativ)

a. Ziel: naturalness

Feste Koeffizienten[a]

ModellItem		Koeffizient	Standard Fehler	t	Sig.	95% Konfidenzintervall		Exp(Coefficient)	95% Konfidenzintervall für Exp(Coefficient)	
						Unterer Wert	Oberer Wert		Unterer Wert	Oberer Wert
Schwellenwert für naturalness=	1	-11,311	2,2844	-4,951	0	-15,799	-6,823	1,22E-05	1,38E-07	0,001
	2	-9,117	2,0079	-4,541	0	-13,061	-5,172	0	2,13E-06	0,006
	3	-6,826	1,9606	-3,481	0,001	-10,677	-2,974	0,001	2,31E-05	0,051
	4	-3,391	1,8631	-1,82	0,069	-7,051	0,269	0,034	0,001	1,308
complexity=1		0,143	0,4418	0,323	0,747	-0,725	1,011	1,153	0,484	2,747
complexity=2		0[b]	.	.	.	.	.	.	.	.
domain=1		-1,751	0,3759	-4,657	0	-2,489	-1,012	0,174	0,083	0,363
domain=2		0[b]	.	.	.	.	.	.	.	.
sentence_type		0,531	0,3807	1,396	0,163	-0,217	1,279	1,701	0,805	3,594
sentence_type*[complexity=1]		0,05	0,1608	0,314	0,754	-0,265	0,366	1,052	0,767	1,442
sentence_type*[complexity=2]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[domain=1]		0,138	0,1157	1,195	0,233	-0,089	0,365	1,148	0,915	1,441
sentence_type*[domain=2]		0[b]	.	.	.	.	.	.	.	.
age_groups=1		-1,67	0,96	-1,74	0,083	-3,556	0,216	0,188	0,029	1,241
age_groups=2		-3,157	0,7974	-3,96	0	-4,724	-1,591	0,043	0,009	0,204
age_groups=3		-3,518	1,264	-2,783	0,006	-6,001	-1,035	0,03	0,002	0,355
age_groups=4		0[b]	.	.	.	.	.	.	.	.
gender=1		-0,053	1,2409	-0,043	0,966	-2,491	2,384	0,948	0,083	10,853
gender=2		0[b]	.	.	.	.	.	.	.	.
experience_linguistics_kat=1		2,87	1,2054	2,381	0,018	0,502	5,238	17,634	1,652	188,274
experience_linguistics_kat=2		1,189	1,0185	1,167	0,244	-0,812	3,19	3,284	0,444	24,281
experience_linguistics_kat=3		0[b]	.	.	.	.	.	.	.	.
experience_fas_kat=1		-0,26	0,8637	-0,301	0,764	-1,957	1,437	0,771	0,141	4,207
experience_fas_kat=2		2,323	1,0201	2,277	0,023	0,319	4,326	10,201	1,375	75,675
experience_fas_kat=3		0[b]	.	.	.	.	.	.	.	.
experience_enc_kat=1		2,016	1,0154	1,985	0,048	0,021	4,011	7,508	1,021	55,19
experience_enc_kat=2		5,198	1,0573	4,916	0	3,121	7,275	180,909	22,666	1443,918
experience_enc_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_openness_kat=2		-1,113	0,6909	-1,611	0,108	-2,47	0,244	0,329	0,085	1,277
bfi_openness_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_conscientiousness_kat=2		-2,157	0,7251	-2,975	0,003	-3,582	-0,733	0,116	0,028	0,48
bfi_conscientiousness_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_extraversion_kat=2		1,671	0,9183	1,82	0,069	-0,133	3,475	5,318	0,876	32,3
bfi_extraversion_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_agreeableness_kat=2		-1,509	0,7207	-2,093	0,037	-2,924	-0,093	0,221	0,054	0,911
bfi_agreeableness_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_neuroticism_kat=1		-4,486	1,129	-3,974	0	-6,704	-2,268	0,011	0,001	0,103
bfi_neuroticism_kat=2		-7,19	0,9143	-7,864	0	-8,986	-5,394	0,001	0	0,005
bfi_neuroticism_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_competence_kat=2		-1,096	0,5962	-1,839	0,067	-2,267	0,075	0,334	0,104	1,078
ta_competence_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_enthusiasm_kat=1		4,962	1,2723	3,9	0	2,463	7,462	142,939	11,74	1740,32
ta_enthusiasm_kat=2		1,936	0,8335	2,323	0,021	0,298	3,573	6,93	1,348	35,632
ta_enthusiasm_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_positive_attitude_kat=1		2,647	1,524	1,737	0,083	-0,347	5,641	14,116	0,707	281,83
ta_positive_attitude_kat=2		0,707	0,713	0,992	0,322	-0,694	2,108	2,028	0,5	8,229
ta_positive_attitude_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_negative_attitude_kat=2		0,424	0,802	0,529	0,597	-1,151	2	1,529	0,316	7,388
ta_negative_attitude_kat=3		0[b]	.	.	.	.	.	.	.	.
sentence_type*[age_groups=1]		-0,56	0,1594	-3,515	0	-0,874	-0,247	0,571	0,417	0,781
sentence_type*[age_groups=2]		-0,131	0,1552	-0,844	0,399	-0,436	0,174	0,877	0,647	1,19
sentence_type*[age_groups=3]		-0,173	0,1743	-0,995	0,32	-0,516	0,169	0,841	0,597	1,184
sentence_type*[age_groups=4]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[gender=1]		0,188	0,2031	0,926	0,355	-0,211	0,587	1,207	0,81	1,799
sentence_type*[gender=2]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[experience_linguistics_kat=1]		-0,395	0,2652	-1,491	0,136	-0,916	0,125	0,673	0,4	1,134
sentence_type*[experience_linguistics_kat=2]		-0,281	0,2126	-1,322	0,187	-0,699	0,137	0,755	0,497	1,146
sentence_type*[experience_linguistics_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[experience_fas_kat=1]		0,357	0,1768	2,019	0,044	0,01	0,704	1,429	1,01	2,022
sentence_type*[experience_fas_kat=2]		-0,248	0,1768	-1,406	0,16	-0,596	0,099	0,78	0,551	1,104
sentence_type*[experience_fas_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[experience_enc_kat=1]		-0,227	0,1822	-1,246	0,213	-0,585	0,131	0,797	0,557	1,14
sentence_type*[experience_enc_kat=2]		-0,599	0,2848	-2,102	0,036	-1,158	-0,039	0,549	0,314	0,961
sentence_type*[experience_enc_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_openness_kat=2]		0,088	0,1309	0,669	0,504	-0,17	0,345	1,092	0,844	1,412
sentence_type*[bfi_openness_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_conscientiousness_kat=2]		0,04	0,1244	0,32	0,749	-0,205	0,284	1,041	0,815	1,329
sentence_type*[bfi_conscientiousness_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_extraversion_kat=2]		-0,298	0,1313	-2,268	0,024	-0,556	-0,04	0,743	0,574	0,961
sentence_type*[bfi_extraversion_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_agreeableness_kat=2]		0,204	0,1197	1,703	0,089	-0,031	0,439	1,226	0,969	1,551
sentence_type*[bfi_agreeableness_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_neuroticism_kat=1]		-0,242	0,331	-0,731	0,465	-0,892	0,408	0,785	0,41	1,504
sentence_type*[bfi_neuroticism_kat=2]		0,468	0,2259	2,071	0,039	0,024	0,912	1,597	1,024	2,489
sentence_type*[bfi_neuroticism_kat=3]		0[b]	.	.	.	.	.	.	.	.

...

Feste Koeffizienten[a] (continued)

Modellterm		Koeffizient	Standard Fehler	t	Sig.	95% Konfidenzintervall Unterer Wert	95% Konfidenzintervall Oberer Wert	Exp(Coefficient)	95% Konfidenzintervall für Exp(Coefficient) Unterer Wert	95% Konfidenzintervall für Exp(Coefficient) Oberer Wert
Schwellenwert für naturalness=	1	-11,311	2,2844	-4,951	0	-15,799	-6,823	1,22E-05	1,38E-07	0,001
	2	-9,117	2,0079	-4,541	0	-13,061	-5,172	0	2,13E-06	0,006
	3	-6,826	1,9606	-3,481	0,001	-10,677	-2,974	0,001	2,31E-05	0,051
	4	-3,391	1,8631	-1,82	0,069	-7,051	0,269	0,034	0,001	1,308
sentence_type*[bfi_openness_kat=2]		0,088	0,1309	0,669	0,504	-0,17	0,345	1,092	0,844	1,412
sentence_type*[bfi_openness_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_conscientiousness_kat=2]		0,04	0,1244	0,32	0,749	-0,205	0,284	1,041	0,815	1,329
sentence_type*[bfi_conscientiousness_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_extraversion_kat=2]		-0,298	0,1313	-2,268	0,024	-0,556	-0,04	0,743	0,574	0,961
sentence_type*[bfi_extraversion_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_agreeableness_kat=2]		0,204	0,1197	1,703	0,089	-0,031	0,439	1,226	0,969	1,551
sentence_type*[bfi_agreeableness_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[bfi_neuroticism_kat=1]		-0,242	0,331	-0,731	0,465	-0,892	0,408	0,785	0,41	1,504
sentence_type*[bfi_neuroticism_kat=2]		0,468	0,2259	2,071	0,039	0,024	0,912	1,597	1,024	2,489
sentence_type*[bfi_neuroticism_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[ta_competence_kat=2]		-0,414	0,1159	-3,569	0	-0,641	-0,186	0,661	0,527	0,83
sentence_type*[ta_competence_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[ta_enthusiasm_kat=1]		-0,153	0,1968	-0,779	0,436	-0,54	0,233	0,858	0,583	1,263
sentence_type*[ta_enthusiasm_kat=2]		-0,109	0,14	-0,779	0,436	-0,384	0,166	0,897	0,681	1,18
sentence_type*[ta_enthusiasm_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[ta_positive_attitude_kat=1]		-0,438	0,2236	-1,961	0,05	-0,878	0,001	0,645	0,416	1,001
sentence_type*[ta_positive_attitude_kat=2]		-0,09	0,1407	-0,642	0,521	-0,367	0,186	0,914	0,693	1,205
sentence_type*[ta_positive_attitude_kat=3]		0[b]	.	.	.	.	.	.	.	.
sentence_type*[ta_negative_attitude_kat=2]		-0,332	0,1191	-2,785	0,006	-0,566	-0,098	0,718	0,568	0,907
sentence_type*[ta_negative_attitude_kat=3]		0[b]	.	.	.	.	.	.	.	.

Wahrscheinlichkeitsverteilung: Multinomial

Verknüpfungsfunktion: Logit (kumulativ)

a. Ziel: naturalness

b. Dieser Koeffizient wurde auf den Wert null gesetzt, da er redundant ist.

A.3.5.2 Recoding of Parameters

For the purpose of better interpretability, the values of metric and ordinal variables were recoded into a maximum of three subgroups. Concerning the Big Five traits and Technical Affinity components, mean values were computed from the respective questionnaire items and assigned to a three-part division of the 5-level Likert scale (low < 1.67, mid < 3.33, high > 3.33). The assessment scale was employed for this purpose, in order to account for the different degrees of a trait manifestation and the interplay of traits as a contiguous characteristic of human personality. Since prior experiences were surveyed as independent competencies to be considered individually, the participants' self-assessments were inspected for each prior knowledge parameter in order to allow a balanced assignment of the participants as possible. As for the four age groups, the included range is shown in Table A.3 below.

Table A.3: The number of participants within each parameter level (range).

Parameter	**Participants per Level**		
	low	mid	high
Age	15 (18–29), 10 (30–44), 6 (45–59), 5 (60–70)		
Experience COP	29 (1)	4 (2)	16 (3)
Experience DAS	14 (1–2)	6 (3)	16 (4–5)
Experience Linguistics	20 (1–3)	12 (4)	4 (5)
Openness	–	11	25
Conscientiousness	–	6	30
Extraversion	–	7	29
Agreeableness	–	6	30
Neuroticism	4	31	1
Competence	–	23	13
Enthusiasm	3	7	26
Positive Attitude	1	34	1
Negative Attitude	–	5	31

A.4 Second User Study Specifying the Influence of Syntax

This appendix section contains the materials used in the second user study specifying the influence of syntactic forms in in-vehicle voice output. First, the preliminary questionnaire (Appendix A.4.1) and post survey (Appendix A.4.2) are presented, followed by the employed syntactic paraphrases (Appendix A.4.3). Second, the explanation of the study content (Appendix A.4.4), procedure (Appendix A.4.5), and the evaluation scale the participants were introduced to (Appendix A.4.6) are provided. Appendix A.4.7 then gives an overview of the dialog tasks that were presented to the participants during the driving simulation. Finally, the results of the evaluation are provided (Appendix A.4.8).

A.4.1 Pre-Survey

The purpose of this questionnaire was to measure individual user characteristics. Besides general demographic and experience-related questions, this survey includes the question-

naire by Karrer *et al.* (2009) concerning a participant's technical affinity (s. "Technikbezogene Selbsteinschätzung") as well as the questionnaire by Rammstedt and Danner (2016) to measure Big Five Personality traits (s. "Persönlichkeitsbezogene Selbsteinschätzung").

Versuchspersonennummer (siehe Terminplan Probandenmanagement)

VP

Versuchsleiter

[Bitte auswählen]

Demographische Daten

Alter

Jahre

Geschlecht

- männlich
- weiblich

Höchster erreichter Bildungsabschluss

- Volksschul-/Hauptschulabschluss
- Mittlere Reife
- Abitur/Fachabitur
- Abgeschlossene Berufsausbildung
- Hochschul-/Fachhochschulabschluss
- Promotion
- anderer:

Aktuelle Tätigkeit

- nicht arbeitstätig
- angestellt
- selbstständig
- in Ausbildung (Schule, Studium, etc.)
- Hausfrau/Hausmann
- Rentner/Pensionär
- andere Tätigkeit

Nutzung und Kenntnisse in der Bedienung von Sprachassistenten

Wie bewerten Sie selbst Ihre Kenntnisse zur Theorie der Sprache (Grammatik, Satzstellung, etc.)?	sehr gering	○○○○○	sehr hoch
Wie bewerten Sie selbst Ihre Vorerfahrung mit Fahrassistenzsystemen?	sehr gering	○○○○○	sehr hoch
Wie bewerten Sie selbst Ihre Vorerfahrung mit ENERGIZING Comfort (S-Klasse Massageprogramme)?	sehr gering	○○○○○	sehr hoch

				nicht beurteilbar
Wie zufrieden sind Sie mit Ihrer Sprachbedienung?	gar nicht zufrieden	○○○○○	sehr zufrieden	○

				Ich habe keine
Wie häufig nutzen Sie die Sprachsteuerung?	nie	○○○○○	sehr häufig	○

Technikbezogene Selbsteinschätzung

Inwieweit treffen die folgenden Aussagen auf Sie persönlich zu?
Lesen Sie bitte jede Aussage aufmerksam durch und entscheiden Sie dann, wie sehr Sie der jeweiligen Aussage zustimmen. Bitte beantworten Sie jede Aussage spontan und wahrheitsgemäß.

Unter dem Begriff „elektronische Geräte" verstehen wir hier Geräte, wie Computer, Handys, Digitalkameras, Geldautomaten, sowie neue Systeme im Auto wie Navigationssysteme. Nicht gemeint sind Werkzeuge (Bohrmaschine, Hammer), Haushaltsgeräte (Toaster, Wasserkocher), Fahrzeuge oder Fahrzeugmotoren.

	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
Ich informiere mich über elektronische Geräte, auch wenn ich keine Kaufabsicht habe.	○	○	○	○	○
Es macht mir Spaß, ein elektronisches Gerät auszuprobieren.	○	○	○	○	○
Elektronische Geräte machen krank.	○	○	○	○	○
Es fällt mir leicht, die Bedienung eines elektronischen Geräts zu lernen.	○	○	○	○	○
Ich kenne mich im Bereich elektronischer Geräte aus.	○	○	○	○	○
Ich bin begeistert, wenn ein neues elektronisches Gerät auf den Markt kommt.	○	○	○	○	○
Elektronische Geräte ermöglichen einen hohen Lebensstandard.	○	○	○	○	○
Ich kenne die meisten Funktionen der elektronischen Geräte, die ich besitze.	○	○	○	○	○
Elektronische Geräte erleichtern mir den Alltag.	○	○	○	○	○
Elektronische Geräte machen unabhängig.	○	○	○	○	○
Elektronische Geräte führen zu geistiger Verarmung.	○	○	○	○	○
Ich liebe es, neue elektronische Geräte zu besitzen.	○	○	○	○	○
Elektronische Geräte verursachen Stress.	○	○	○	○	○
Elektronische Geräte erhöhen die Sicherheit.	○	○	○	○	○
Ich habe bzw. hätte Verständnisprobleme beim Lesen von Elektronik- Computerzeitschriften.	○	○	○	○	○
Elektronische Geräte helfen, an Informationen zu gelangen.	○	○	○	○	○
Elektronische Geräte machen vieles umständlicher.	○	○	○	○	○
Elektronische Geräte verringern den persönlichen Kontakt zwischen den Menschen.	○	○	○	○	○
Ich gehe gern in den Fachhandel für elektronische Geräte.	○	○	○	○	○

Persönlichkeitsbezogene Selbsteinschätzung

Inwieweit treffen die folgenden Aussagen auf Sie persönlich zu? Lesen Sie bitte jede Aussage aufmerksam durch und entscheiden Sie dann, wie sehr Sie der jeweiligen Aussage zustimmen. Bitte beantworten Sie jede Aussage spontan und wahrheitsgemäß.

Ich...	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
... bin gesprächig und unterhalte mich gern.	○	○	○	○	○
... neige dazu, andere zu kritisieren.	○	○	○	○	○
... erledige Aufgaben gründlich.	○	○	○	○	○
... bin deprimiert, niedergeschlagen.	○	○	○	○	○
... bin originell, entwickle neue Ideen.	○	○	○	○	○
... bin eher zurückhaltend, reserviert.	○	○	○	○	○
... bin hilfsbereit und selbstlos gegenüber anderen.	○	○	○	○	○
... bin manchmal unsorgfältig und schluderig.	○	○	○	○	○
... bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	○	○	○	○	○

Ich...	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
... bin vielseitig interessiert.	○	○	○	○	○
... bin voller Energie und Tatendrang.	○	○	○	○	○
... bin häufig in Streitereien verwickelt.	○	○	○	○	○
... arbeite zuverlässig und gewissenhaft.	○	○	○	○	○
... reagiere leicht angespannt.	○	○	○	○	○
... bin tiefsinnig, denke gerne über Sachen nach.	○	○	○	○	○
... bin begeisterungsfähig und kann andere leicht mitreißen.	○	○	○	○	○
... bin nicht nachtragend, vergebe anderen leicht.	○	○	○	○	○
... bin eher unordentlich.	○	○	○	○	○

Ich...	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
... mache mir viele Sorgen.	○	○	○	○	○
... habe eine aktive Vorstellungskraft, bin fantasievoll.	○	○	○	○	○
... bin eher der „stille Typ", wortkarg.	○	○	○	○	○
... schenke anderen Vertrauen, glaube an das Gute im Menschen.	○	○	○	○	○
... bin bequem, neige zur Faulheit.	○	○	○	○	○
... bin emotional ausgeglichen, nicht leicht aus der Fassung zu bringen.	○	○	○	○	○
... bin erfinderisch und einfallsreich.	○	○	○	○	○
... bin durchsetzungsfähig, energisch.	○	○	○	○	○
... kann mich kalt und distanziert verhalten.	○	○	○	○	○

Ich...	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
... harre aus (und arbeite weiter), bis die Aufgabe fertig ist.	○	○	○	○	○
... kann launisch sein, habe schwankende Stimmungen.	○	○	○	○	○
... schätze künstlerische und ästhetische Eindrücke.	○	○	○	○	○
... bin manchmal schüchtern und gehemmt.	○	○	○	○	○
... bin rücksichtsvoll zu anderen, einfühlsam.	○	○	○	○	○
... bin tüchtig und arbeite flott.	○	○	○	○	○
... bleibe ruhig, selbst in Stresssituationen.	○	○	○	○	○
... mag es, wenn Aufgaben routinemäßig zu erledigen sind.	○	○	○	○	○
... gehe aus mir heraus, bin gesellig.	○	○	○	○	○

Ich...	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
... kann mich schroff und abweisend anderen gegenüber verhalten.	○	○	○	○	○
... mache Pläne und führe sie auch durch.	○	○	○	○	○
... werde leicht nervös und unsicher.	○	○	○	○	○
... stelle gerne Überlegungen an, spiele mit abstrakten Ideen.	○	○	○	○	○
... habe nur wenig künstlerisches Interesse.	○	○	○	○	○
... verhalte mich kooperativ, ziehe Zusammenarbeit dem Wettbewerb vor.	○	○	○	○	○
... bin leicht ablenkbar, bleibe nicht bei der Sache.	○	○	○	○	○
... kenne mich gut in Musik, Kunst oder Literatur aus.	○	○	○	○	○
... habe oft Krach mit anderen.	○	○	○	○	○

In welcher körperlichen Verfassung sind Sie im Moment?

○ Ich befinde mich in meinem üblichen Fitness- & Gesundheitszustand.

○ Ich bin derzeit in einer schlechteren Verfassung als üblich (z.B. Erkältung).

Bitte geben Sie auf einer Skala von 0 – 20 an, wie Sie sich zu diesem Zeitpunkt fühlen.

Der Wert **„0"** bedeutet dabei, dass Sie **keinerlei Beschwerden** haben und es Ihnen **sehr gut** geht, während **„20"** heißt, dass Sie sich **extrem unwohl** fühlen. Bei dieser Bewertung geht es in erster Linie um eine eventuelle Übelkeit.

Bitte Wert eintragen: [ ]

Vielen Dank!

Ihre Antworten wurden gespeichert, Sie können das Browser-Fenster nun schließen.

A.4.2 Intermediate and Post Survey

The purpose of this questionnaire was to measure individual driver distraction and and consists of the DALI questionnaire based on Hofmann (2015).

Versuchspersonennummer (siehe Terminplan Probandenmanagement)

VP

[Bitte auswählen]

[Bitte auswählen]

Bewertung Ihrer Erfahrung mit dem System

*Bitte beziehen Sie Ihre Antwort immer auf die **Gesamtanforderungen**, d. h. die Anforderungen, die Sie während Ihrer Fahrt **in Verbindung** mit dem Frage-Antwort-Dialog **und** der anschließenden Bewertung erlebt haben. Bitte beantworten Sie jede Aussage möglichst spontan.*

Wie hoch waren die Anforderungen an die globale Aufmerksamkeit?

Erklärung: Insgesamt alle mentalen – denken, entscheiden, ... - visuellen und auditiven Faktoren, die insgesamt während des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen.

sehr gering sehr hoch

Wie hoch waren die visuellen Anforderungen?

Erklärung: Visuelle Faktoren, die während des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen – alles, was mit dem Sehen zu tun hat.

sehr gering sehr hoch

Wie hoch waren die auditiven Anforderungen?

Erklärung: Auditive Faktoren, die während des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen – alles, was mit Gehörtem zu tun hat.

sehr gering sehr hoch

Wie stark war das Stressniveau?

Erklärung: Stressniveau während des Autofahrens mit Frage-Antwort-Dialogen wie Irritation, Müdigkeit, Unsicherheit, Entmutigung, etc.

sehr gering sehr hoch

Wie hoch war die zeitliche Anforderung?

Erklärung: Gefühlte Belastung und spezifische Beeinträchtigung durch die schnelle Abfolge der Aufgaben.

sehr gering sehr hoch

Wie stark war der Interferenzfaktor?

Erklärung: Die Fahrerbeanspruchung und ihre Auswirkung auf die Fahrleistung durch den gleichzeitig ablaufenden Frage-Antwort-Dialog während des Autofahrens.

sehr gering sehr hoch

A.4.3 Syntactic Paraphrases

This appendix section provides an overview of the syntactic paraphrases for the two domains DAS and COP and the three question types What, How and When. They were generated in the form of both sentence types MCV and RCV and employed in this driving simulation study as explanatory voice prompts as respective answers to the questions “What is...?,” “How does ... work?,” and “When can I use ...?.”

F	Q	MCV	RCV
Behaglichkeit (“Well-being”)	What	Das Programm Behaglichkeit kann Ihre körperliche und mentale Entspannung unterstützen. Es kombiniert eine Hotstone-Rückenmassage mit lokaler Wärme.	Das Programm Behaglichkeit, das Ihre körperliche und mentale Entspannung unterstützen kann, kombiniert eine Hotstone-Rückenmassage mit lokaler Wärme.
	How	Das Programm Behaglichkeit kann Sie durch eine Rückenmassage entspannen. Es nutzt zusätzlich entspannende Musik und eine lila Ausleuchtung des Innenraums.	Das Programm Behaglichkeit, das Sie durch eine Rückenmassage entspannen kann, nutzt zusätzlich entspannende Musik und eine lila Ausleuchtung des Innenraums.
	When	Das Programm Behaglichkeit kann Ihnen in angespannten Fahrsituationen zu Ihrer Entspannung dienen. Es steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.	Das Programm Behaglichkeit, das Ihnen in angespannten Fahrsituationen zu Ihrer Entspannung dienen kann, steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.
Vergügen (“Joy”)	What	Das Programm Vergnügen kann Ihnen in ermüdenden Fahrsituationen für eine positive Stimmung dienen. Es steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.	Das Programm Vergnügen, das Ihnen in ermüdenden Fahrsituationen für eine positive Stimmung dienen kann, steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.
	How	Das Programm Vergnügen kann Sie durch ein wohltuendes Massageprogramm entspannen. Es nutzt dabei zusätzlich mäßig schnelle Musik und eine gelbe Ausleuchtung des Innenraums.	Das Programm Vergnügen, das Sie durch ein wohltuendes Massageprogramm entspannen kann, nutzt dabei zusätzlich mäßig schnelle Musik und eine gelbe Ausleuchtung des Innenraums.

F	Q	MCV	RCV
	When	Das Programm Vergnügen kann eine positive Stimmung begünstigen. Es kombiniert ein wohltuendes Massageprogramm mit mäßig schneller Musik.	Das Programm Vergnügen, das eine positive Stimmung begünstigen kann, kombiniert ein wohltuendes Massageprogramm mit mäßig schneller Musik.

F	Q	MCV	RCV
Vitalität (“Vitality”)	What	Das Programm Vitalität kann Ihrer Ermüdung während der Fahrt entgegenwirken. Es kombiniert aktivierende Musik mit einer belebenden Massage.	Das Programm Vitalität, das Ihrer Ermüdung während der Fahrt entgegenwirken kann, kombiniert aktivierende Musik mit einer belebenden Massage.
	How	Das Programm Vitalität kann Sie durch eine aktivierende Musik stimulieren. Es nutzt zusätzlich eine belebende Massage und eine rote Ausleuchtung des Innenraums.	Das Programm Vitalität, das Sie durch eine aktivierende Musik stimulieren kann, nutzt zusätzlich eine belebende Massage und eine rote Ausleuchtung des Innenraums.
	When	Das Programm Vitalität kann Ihnen in monotonen Fahrsituationen für eine aktivierende Stimulation dienen. Es steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.	Das Programm Vitalität, das Ihnen in monotonen Fahrsituationen für eine aktivierende Stimulation dienen kann, steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.
Wärme (“Warmth”)	What	Das Programm Wärme kann Ihr Wohlbefinden steigern. Es kombiniert die Beheizung von Lenkrad und Sitzen mit einer warmen Ausleuchtung des Innenraums.	Das Programm Wärme, das Ihr Wohlbefinden steigern kann, kombiniert die Beheizung von Lenkrad und Sitzen mit einer warmen Ausleuchtung des Innenraums.
	How	Das Programm Wärme kann Sie gezielt durch eine wohlige Wärme in Sitz und Lenkrad entspannen. Es nutzt zusätzlich die Ausleuchtung des Innenraums in einem warmen Orange.	Das Programm Wärme, das Sie gezielt durch eine wohlige Wärme in Sitz und Lenkrad entspannen kann, nutzt zusätzlich die Ausleuchtung des Innenraums in einem warmen Orange.
	When	Das Programm Wärme kann Ihnen in belastenden Fahrsituationen für ein gemütliches Ambiente dienen. Es steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.	Das Programm Wärme, das Ihnen in belastenden Fahrsituationen für ein gemütliches Ambiente dienen kann, steht ab fünf Minuten nach Start des Multimediasystems zur Verfügung.

F	Q	MCV	RCV
Abstands-Assistent (“Space Assist”)	What	Der aktive Abstands-Assistent kann Sie auf langen Strecken und im Stop-and-Go-Verkehr unterstützen. Er ist bis zu einer Geschwindigkeit von 210 km/h einsetzbar.	Der aktive Abstands-Assistent, der Sie auf langen Strecken und im Stop-and-Go-Verkehr unterstützen kann, ist bis zu einer Geschwindigkeit von 210 km/h einsetzbar.
	How	Der aktive Abstands-Assistent warnt Sie optisch und akustisch. Bei einem zu geringen Abstand zu Ihrem vorausfahrenden Fahrzeug bremst er selbst ab.	Der aktive Abstands-Assistent, der Sie optisch und akustisch warnt, bremst bei einem zu geringen Abstand zu Ihrem vorausfahrenden Fahrzeug selbst ab.
	When	Der aktive Abstands-Assistent kann einen sicheren Abstand zum vorausfahrenden Fahrzeug halten. So verringert er das Risiko von Auffahrunfällen.	Der aktive Abstands-Assistent, der einen sicheren Abstand zum vorausfahrenden Fahrzeug halten kann, verringert so das Risiko von Auffahrunfällen.
Brems-Assistent (“Brake Assist”)	What	Der aktive Brems-Assistent kann Sie bei einer Kollisionsgefahr mit Fahrzeugen oder Fußgängern unterstützen. Er steht Ihnen bis zu einer Geschwindigkeit bis 250 km/h zur Verfügung.	Der aktive Brems-Assistent, der Sie bei einer Kollisionsgefahr mit Fahrzeugen oder Fußgängern unterstützen kann, steht Ihnen bis zu einer Geschwindigkeit von 250 km/h zur Verfügung.
	How	Der aktive Brems-Assistent warnt Sie zuerst akustisch. In kritischen Situationen löst er dann eine autonome Bremsung aus, notfalls bis zu einer Vollbremsung.	Der aktive Brems-Assistent, der Sie zuerst akustisch warnt, löst dann in kritischen Situationen eine autonome Bremsung aus, notfalls bis zu einer Vollbremsung.
	When	Der aktive Brems-Assistent kann mit Hilfe der Abstandswarnfunktion das Risiko einer Kollision erkennen. So vermeidet er die Gefahr von Auffahrunfällen.	Der aktive Brems-Assistent, der mit Hilfe der Abstandswarnfunktion das Risiko einer Kollision erkennen kann, vermeidet so die Gefahr von Auffahrunfällen.

F	Q	MCV	RCV
Spurhalte-Assistent (“Lane Keeping Assist”)	What	Der aktive Spurhalte-Assistent kann Sie vor unbeabsichtigten Spurwechseln schützen. So verringert er die Gefahr einer seitlichen Kollision.	Der aktive Spurhalte-Assistent, der Sie vor unbeabsichtigten Spurwechseln schützen kann, verringert so die Gefahr einer seitlichen Kollision.
	How	Der aktive Spurhalte-Assistent warnt Sie zuerst durch eine Vibration des Lenkrads. Bei einem unbeabsichtigten Spurwechsel führt er Ihr Fahrzeug dann eigenständig zurück in die ursprüngliche Spur.	Der aktive Spurhalte-Assistent, der Sie zuerst durch eine Vibration des Lenkrads warnt, führt Ihr Fahrzeug dann bei einem unbeabsichtigten Spurwechsel eigenständig zurück in die ursprüngliche Spur.
	When	Der aktive Spurhalte-Assistent kann Sie sowohl auf Autobahnen als auch im Stadtverkehr unterstützen. Er ist bis zu einer Geschwindigkeit von 210 km/h einsetzbar.	Der aktive Spurhalte-Assistent, der Sie sowohl auf Autobahnen als auch im Stadtverkehr unterstützen kann, ist bis zu einer Geschwindigkeit von 210 km/h einsetzbar.
Totwinkel-Assistent (“Blind Spot Assist”)	What	Der aktive Totwinkel-Assistent warnt Sie durch einen Signalton. Bei einem Fahrzeug im toten Winkel aktiviert er außerdem eine rote Warnleuchte im jeweiligen Außenspiegel.	Der Totwinkel-Assistent kann Fahrzeuge im toten Winkel erkennen. So vermeidet er das Risiko von Kollisionen mit anderen Fahrzeugen.
	How	Der aktive Totwinkel-Assistent, der Sie durch einen Signalton warnt, aktiviert außerdem bei einem Fahrzeug im toten Winkel eine rote Warnleuchte im jeweiligen Außenspiegel.	Der Totwinkel-Assistent, der Fahrzeuge im toten Winkel erkennen kann, vermeidet so das Risiko von Kollisionen mit anderen Fahrzeugen.
	When	Der aktive Totwinkel-Assistent kann Sie im Stadtverkehr, auf Schnellstraßen und auf Autobahnen unterstützen. Er steht Ihnen bis zu einer Geschwindigkeit von 210 km/h zur Verfügung.	Der aktive Totwinkel-Assistent, der Sie im Stadtverkehr, auf Schnellstraßen und auf Autobahnen unterstützen kann, steht Ihnen bis zu einer Geschwindigkeit von 210 km/h zur Verfügung.

A.4.4 Explanation of Study Content

IHRE HEUTIGE FAHRT

Sie haben heute die Möglichkeit einen Sprachassistenten zu erleben und ihn anschließend zu bewerten. Der Sprachassistent liefert Ihnen während der Fahrt zusätzliche Informationen zum Fahrzeug und seinen Funktionen. Uns interessiert hier Ihre ganz persönliche Meinung. Sie leisten damit einen wertvollen Beitrag zur Weiterentwicklung des Systems.

Während einer Fahrt auf der Landstraße lernen Sie das Sprachdialogsystem in zwei Situationen kennen:

- In einer Situation fahren Sie selbst. Sie folgen einem vorausfahrenden Fahrzeug mit möglichst konstanter Geschwindigkeit von 100 km/h und halten bitte einen Abstand von ca. 100 m ein (entspricht zwei Leitpfosten).
- Die andere Situation besteht aus einer hochautomatisierten Fahrt, hier übernimmt das Fahrzeug für Sie das Lenken, Gas geben und Bremsen.

Während Ihrer Fahrt erhalten Sie die Aufgabe, dem Sprachassistenten Fragen aus vorgegebenen Themenbereichen zu stellen (*Fahrerassistenzsysteme* und *Energizing Comfort Programme*). Sie aktivieren den Sprachassistenten, indem Sie „Hallo Mercedes“ sagen und formulieren dann direkt Ihre Frage. Danach bewerten Sie die Antwort des Sprachassistenten im Hinblick auf Verständlichkeit und Natürlichkeit.

Mit ***Verständlichkeit*** ist gemeint:

- **Verstehen Sie die Antwort des Sprachassistenten, also verstehen Sie was gesagt wird?**
- Ist die Antwort intuitiv und sofort verständlich?
- Oder ist das Gesagte erst mit etwas Zeit und Nachdenken zu verstehen?
- Werden die Antworten aus Ihrer Sicht verständlich gestaltet oder bewerten Sie die Formulierungen z.B. als zu simpel oder zu kompliziert?

Mit ***Natürlichkeit*** ist gemeint:

- **Beurteilen Sie das Gespräch, also die Frage-Antwort-Sequenzen mit dem Sprachassistenten als angenehm?**
- Entspricht die Formulierung der Antworten Ihren Erwartungen an ein System?
- Oder wünschen Sie sich, dass ein Sprachassistent mit Ihnen auf eine andere Weise spricht (z.B. einfacher, förmlicher, umgangssprachlicher…)?
- Beurteilen Sie den Sprachstil des Assistenten als angenehm (natürliche Sprache) und hören den Formulierungen gerne zu, wie etwa im Gespräch mit anderen Menschen?

Bitte beachten Sie, dass sich die Bewertungen auf die **Qualität und Gestaltung der Antwort** des Sprachassistenten beziehen und **nicht** darauf, ob…

- Ihnen die Stimme des Sprachassistenten gefällt oder sympathisch ist.
- der Sprachassistent eine fehlerfreie Aussprache hat.

Während der gesamten Fahrt haben Sie Sprechkontakt zu Ihrer Versuchsleiterin. Sollten Sie sich zu irgendeinem Zeitpunkt nicht wohl fühlen, geben Sie bitte sofort Bescheid. Sie können die Fahrt dann zu jedem Zeitpunkt unterbrechen bzw. abbrechen.

Wir wünschen gute Fahrt!

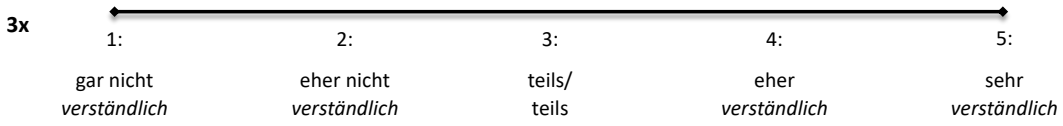
A.4.5 Explanation of the Study Procedure

The graphics have been removed due to copyright limitations.

A.4.6 Evaluation Scale

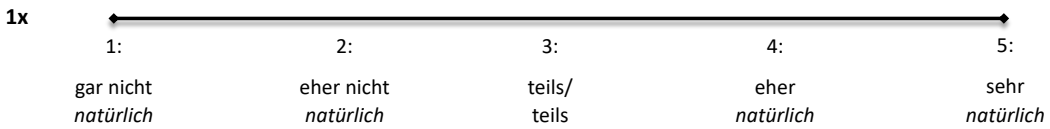
BEWERTUNGSSKALA 1

Auf einer Skala von 1 (gar nicht *verständlich*) bis 5 (sehr *verständlich*), wie bewerten Sie die gehörte Antwort?



BEWERTUNGSSKALA 2

Auf einer Skala von 1 (gar nicht *natürlich*) bis 5 (sehr *natürlich*), wie bewerten Sie die gehörten Antworten zum Nothalt-Assistenten?



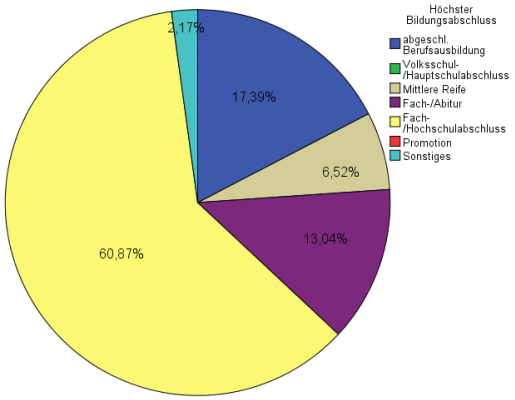
A.4.7 Dialog Tasks

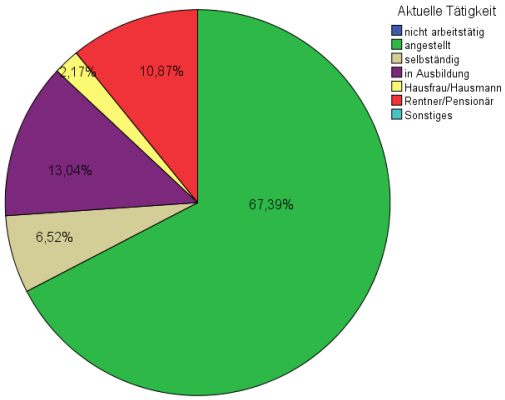
This section contains the dialog tasks employed in the user study. During the driving simulator study, they were projected onto the head-unit to indicate the question the participants should formulate. Based on the *Nothalt-Assistent* (“Emergency Stop Assist”), the participants were explained that each dialog task consisted of a picture and name representing the vehicle function to be inquired and the question type What, How or When.

The graphics have been removed due to copyright limitations.

A.4.8 Results of the second User Study on the Influence of Syntax

This section provides an overview of the results obtained for the second user study. The following figures contain supplementary results from the pre-survey: The majority of participants reported having a university degree and being in a salaried or apprenticeship position at the time of the study.





A.4.8.1 Recoding of Parameters

Similar to the first user study (Appendix A.3.5.2), the values of metric and ordinal variables were recoded into a maximum of three subgroups for the purpose of better interpretability. Since human personality following the Big Five model is considered to be manifested by an interplay of the individual traits in this work, the assessment scale was employed as a common measure to account for the different degrees of trait manifestations. For this purpose, mean values were computed from the respective Big Five and Technical Affinity questionnaire items

Table A.5: The number of participants within each parameter level.

Parameter	**Participants per Level**		
	low	mid	high
Age	14 (18–29), 11 (30–44), 14 (45–59), 7 (60–70)		
Experience COP	34 (1–2)	9 (3)	3 (4–5)
Experience DAS	14 (1–2)	14 (3)	17 (4–5)
Experience Linguistics	2 (1–2)	11 (3)	33 (4–5)
Agreeableness	–	11	35
Conscientiousness	–	6	40
Extraversion	–	21	25
Neuroticism	24	20	2
Openness	–	28	18
Competence	–	20	26
Enthusiasm	3	13	30
Positive Attitude	–	14	32
Negative Attitude	–	6	40

and assigned to a three-part division of the 5-level Likert scale (low $<$ 1.67, mid $<$ 3.33, high $>$ 3.33). In contrast, prior experiences were interpreted as independent competencies to be considered individually. Therefore, the participants' self-assessments were inspected for each prior knowledge parameter in order to allow a balanced assignment of the participants. As for the four age groups, the included range is shown in Table A.5.

A.4.8.2 Generalized Linear Mixed Models

In the following, the results of the explorative generalized linear mixed models are presented.

Dependent Variable Naturalness

Modellübersicht

Ziel		naturalness
Messniveau		Ordinal
Wahrscheinlichkeitsverteilung		Multinomial
Verknüpfungsfunktion		Logit (kumulativ)
Informationskriterium	Akaike (korrigiert)	7043,899
	Bayes	7047,648

Informationskriterien beruhen auf der -2 Log-Likelihood (7041,886) und dienen zum Modellvergleich. Modelle mit kleineren Werten für Informationskriterien weisen eine bessere Anpassung auf.

Zusammenfassung der Fallverarbeitung

	N	Prozent
Eingeschlossen	368	100,00%
Ausgeschlossen	0	0,00%
Gesamtergebnis	368	100,00%

Klassifikation

Gesamtprozent korrekt = 80,4%[a]

		Vorhergesagt			
Beobachtet		eher unnatürlich	teils/teils	eher natürlich	sehr natürlich
eher unnatürlich	Anzahl	0	7	0	0
	% in 'Beobachtet'	0,00%	100,00%	0,00%	0,00%
teils/teils	Anzahl	0	47	18	0
	% in 'Beobachtet'	0,00%	72,30%	27,70%	0,00%
eher natürlich	Anzahl	0	10	116	21
	% in 'Beobachtet'	0,00%	6,80%	78,90%	14,30%
sehr natürlich	Anzahl	0	0	16	133
	% in 'Beobachtet'	0,00%	0,00%	10,70%	89,30%

a. Ziel: naturalness

Kovarianzparameter-Übersicht

Kovarianzparameter	Residualeffekt	0
	Zufällige Effekte	1
Design-Matrix-Spalten	Feste Effekte	122
	Zufällige Effekte	1[a]
Gemeinsame Subjekte		46

Gemeinsame Subjekte beruhen auf den Subjektspezifikationen für den Residualeffekt und die zufälligen Effekte und dienen dazu, die Daten aufzuteilen, um eine bessere Leistungsfähigkeit zu erreichen.

a. Dies ist die Anzahl an Spalten pro gemeinsamem Subjekt.

Block für zufällige Effekte 1

Block für zufällige Effekte	Konstanter Term
Konstanter Term	4,452

Kovarianzstruktur: Skalierte Identität

Subjektspezifikation: ID

Zufälliger Effekt

					95% Konfidenzintervall	
Zufälliger Effekt Kovarianz	Schätzer	Standard Fehler	Z	Sig.	Unterer Wert	Oberer Wert
Varianz	4,452	1,676	2,657	0,008	2,129	9,309

Kovarianzstruktur: Skalierte Identität

Subjektspezifikation: ID

Feste Effekte[a]

Quelle	F	df1	df2	Sig.
Korrigiertes Modell	364231,928	45	318	0,000
complexity	0,039	1	318	0,843
domain	0,074	1	318	0,786
sentence_type	63,373	1	318	0,000
complexity * sentence_type	0,629	1	318	0,428
domain * sentence_type	0,098	1	318	0,755
age_groups	11,2	3	318	0,000
gender	71,89	1	318	0,000
experience_linguistics_kat	17,284	2	318	0,000
experience_fas_kat	1,865	2	318	0,157
experience_enc_kat	0,15	2	318	0,861
bfi_openness_kat	12,098	1	318	0,001
bfi_conscientiousness_kat	5,409	1	318	0,021
bfi_extraversion_kat	3,773	1	318	0,053
bfi_agreeableness_kat	11,076	1	318	0,001
bfi_neuroticism_kat	7,158	2	318	0,001
ta_competence_kat	0,001	1	318	0,976
ta_enthusiasm_kat	9,155	2	318	0,000
ta_positive_attitude_kat	1,399	1	318	0,238
ta_negative_attitude_kat	2,287	1	318	0,131
age_groups * sentence_type	1,601	3	318	0,189
gender * sentence_type	0,033	1	318	0,855
experience_linguistics_kat * sentence_type	93,287	2	318	0,000
experience_fas_kat * sentence_type	1,218	2	318	0,297
experience_enc_kat * sentence_type	0,101	2	318	0,904
bfi_openness_kat * sentence_type	3,353	1	318	0,068
bfi_conscientiousness_kat * sentence_type	5,093	1	318	0,025
bfi_extraversion_kat * sentence_type	1,699	1	318	0,193
bfi_agreeableness_kat * sentence_type	6,3	1	318	0,013
bfi_neuroticism_kat * sentence_type	3,781	2	318	0,024
ta_competence_kat * sentence_type	0,033	1	318	0,856
ta_enthusiasm_kat * sentence_type	0,188	2	318	0,829
ta_positive_attitude_kat * sentence_type	0,532	1	318	0,466
ta_negative_attitude_kat * sentence_type	0,324	1	318	0,570

Wahrscheinlichkeitsverteilung: Multinomial

Verknüpfungsfunktion: Logit (kumulativ)[a]

a. Ziel: naturalness

Feste Koeffizienten[a]

Modellterm		Koeffizient	Standard Fehler	t	Sig.	95% Konfidenzintervall			95% Konfidenzintervall für Exp(Coefficient)	
						Unterer Wert	Oberer Wert	Exp(Coefficient)	Unterer Wert	Oberer Wert
Schwellenwert für naturalness=	2	-12,677	2,6498	-4,784	0	-17,89	-7,464	3,12E-06	1,70E-08	0,001
	3	-8,716	2,6113	-3,338	0,001	-13,853	-3,578	0	9,63E-07	0,028
	4	-4,121	2,3407	-1,761	0,079	-8,727	0,484	0,016	0,00E+00	1,622
complexity=1		0,113	0,3923	0,289	0,773	-0,659	0,885	1,12	0,518	2,423
complexity=2		0[b]	.	.	.	.	.	.	.	.
domain=1		0,014	0,422	0,033	0,974	-0,816	0,844	1,014	0,442	2,326
domain=2		0[b]	.	.	.	.	.	.	.	.
sentence_type=1		-3,172	2,5245	-1,257	0,21	-8,139	1,794	0,042	0	6,016
sentence_type=2		0[b]	.	.	.	.	.	.	.	.
[complexity=1]*[sentence_type=1]		-0,36	0,4537	-0,793	0,428	-1,252	0,533	0,698	0,286	1,704
[complexity=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[complexity=2]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[complexity=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[domain=1]*[sentence_type=1]		0,148	0,4733	0,312	0,755	-0,783	1,079	1,159	0,457	2,942
[domain=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[domain=2]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[domain=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
age_groups=1		-5,774	1,4283	-4,043	0	-8,584	-2,964	0,003	0	0,052
age_groups=2		-4,971	1,3915	-3,573	0	-7,709	-2,234	0,007	0	0,107
age_groups=3		-0,98	1,0687	-0,917	0,36	-3,082	1,123	0,375	0,046	3,073
age_groups=4		0[b]	.	.	.	.	.	.	.	.
gender=1		-5,03	0,65	-7,738	0	-6,309	-3,751	0,007	0,002	0,023
gender=2		0[b]	.	.	.	.	.	.	.	.
experience_linguistics_kat=1		-13,637	1,8923	-7,207	0	-17,36	-9,914	0,000001195	2,888E-08	0,00004948
experience_linguistics_kat=2		-3,456	0,5725	-6,037	0	-4,582	-2,33	0,032	0,01	0,097
experience_linguistics_kat=3		0[b]	.	.	.	.	.	.	.	.
experience_fas_kat=1		-1,142	0,9964	-1,146	0,253	-3,102	0,819	0,319	0,045	2,268
experience_fas_kat=2		1,13	0,8611	1,312	0,191	-0,565	2,824	3,094	0,569	16,841
experience_fas_kat=3		0[b]	.	.	.	.	.	.	.	.
experience_enc_kat=1		-0,813	1,2972	-0,627	0,531	-3,365	1,739	0,443	0,035	5,691
experience_enc_kat=2		-0,508	1,185	-0,429	0,668	-2,84	1,823	0,602	0,058	6,191
experience_enc_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_openness_kat=2		1,571	0,6041	2,6	0,01	0,382	2,76	4,812	1,466	15,795
bfi_openness_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_conscientiousness_kat=2		-3,975	1,3471	-2,951	0,003	-6,625	-1,324	0,019	0,001	0,266
bfi_conscientiousness_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_extraversion_kat=2		-1,41	1,0052	-1,403	0,162	-3,387	0,568	0,244	0,034	1,764
bfi_extraversion_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_agreeableness_kat=2		2,125	1,052	2,02	0,044	0,055	4,194	8,37	1,056	66,312
bfi_agreeableness_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_neuroticism_kat=1		2,441	1,5851	1,54	0,124	-0,677	5,56	11,49	0,508	259,863
bfi_neuroticism_kat=2		2,615	1,2374	2,113	0,035	0,18	5,05	13,668	1,198	155,967
bfi_neuroticism_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_competence_kat=2		-0,054	1,1512	-0,047	0,962	-2,319	2,21	0,947	0,098	9,12
ta_competence_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_enthusiasm_kat=1		-4,638	1,5294	-3,033	0,003	-7,648	-1,629	0,01	0	0,196
ta_enthusiasm_kat=2		-0,008	1,1368	-0,007	0,994	-2,245	2,229	0,992	0,106	9,286
ta_enthusiasm_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_positive_attitude_kat=2		0,719	0,7659	0,939	0,349	-0,788	2,226	2,052	0,455	9,262
ta_positive_attitude_kat=3		0[b]	.	.	.	.	.	.	.	.
ta_negative_attitude_kat=2		-2,026	1,4069	-1,44	0,151	-4,794	0,742	0,132	0,008	2,101
ta_negative_attitude_kat=3		0[b]	.	.	.	.	.	.	.	.
[age_groups=1]*[sentence_type=1]		0,639	1,0848	0,589	0,556	-1,496	2,773	1,894	0,224	16,006
[age_groups=2]*[sentence_type=1]		1,218	1,0215	1,192	0,234	-0,792	3,228	3,381	0,453	25,229
[age_groups=3]*[sentence_type=1]		-0,05	1,1345	-0,044	0,965	-2,282	2,182	0,951	0,102	8,868
[age_groups=4]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[age_groups=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[age_groups=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[age_groups=3]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[age_groups=4]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[gender=1]*[sentence_type=1]		0,093	0,5114	0,183	0,855	-0,913	1,099	1,098	0,401	3,002
[gender=2]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[gender=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[gender=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=1]*[sentence_type=1]		22,291	1,6323	13,656	0	19,079	25,502	4795322603	193224878,4	1,19007E+11
[experience_linguistics_kat=2]*[sentence_type=1]		0,406	0,3974	1,023	0,307	-0,375	1,188	1,501	0,687	3,281
[experience_linguistics_kat=3]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=3]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=1]*[sentence_type=1]		-0,663	0,6521	-1,016	0,31	-1,946	0,62	0,515	0,143	1,86
[experience_fas_kat=2]*[sentence_type=1]		-0,858	0,6623	-1,295	0,196	-2,161	0,445	0,424	0,115	1,561
[experience_fas_kat=3]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=3]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.

...

(continued)

[experience_enc_kat=1]*[sentence_type=1]	0,268	1,4156	0,189	0,85	-2,517	3,053	1,307	0,081	21,173
[experience_enc_kat=2]*[sentence_type=1]	0,101	1,4987	0,067	0,947	-2,848	3,049	1,106	0,058	21,096
[experience_enc_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[experience_enc_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_enc_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_enc_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_openness_kat=2]*[sentence_type=1]	1,006	0,5495	1,831	0,068	-0,075	2,087	2,735	0,928	8,062
[bfi_openness_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_openness_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_openness_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_conscientiousness_kat=2]*[sentence_type=1]	2,265	1,0037	2,257	0,025	0,29	4,24	9,631	1,337	69,387
[bfi_conscientiousness_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_conscientiousness_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_conscientiousness_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_extraversion_kat=2]*[sentence_type=1]	-0,671	0,5151	-1,303	0,193	-1,685	0,342	0,511	0,185	1,408
[bfi_extraversion_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_extraversion_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_extraversion_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_agreeableness_kat=2]*[sentence_type=1]	2,749	1,0955	2,51	0,013	0,594	4,905	15,635	1,812	134,929
[bfi_agreeableness_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_agreeableness_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_agreeableness_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=1]*[sentence_type=1]	3,355	1,2856	2,61	0,009	0,826	5,884	28,642	2,283	359,338
[bfi_neuroticism_kat=2]*[sentence_type=1]	2,986	1,1241	2,656	0,008	0,774	5,197	19,797	2,168	180,752
[bfi_neuroticism_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_competence_kat=2]*[sentence_type=1]	0,166	0,9138	0,182	0,856	-1,632	1,964	1,181	0,196	7,127
[ta_competence_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_competence_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_competence_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=1]*[sentence_type=1]	-0,028	1,2612	-0,022	0,982	-2,509	2,454	0,973	0,081	11,629
[ta_enthusiasm_kat=2]*[sentence_type=1]	-0,431	0,9135	-0,472	0,637	-2,228	1,366	0,65	0,108	3,92
[ta_enthusiasm_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_positive_attitude_kat=2]*[sentence_type=1]	0,393	0,5395	0,729	0,466	-0,668	1,455	1,482	0,513	4,283
[ta_positive_attitude_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_positive_attitude_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_positive_attitude_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_negative_attitude_kat=2]*[sentence_type=1]	0,467	0,8212	0,569	0,57	-1,148	2,083	1,596	0,317	8,03
[ta_negative_attitude_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_negative_attitude_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_negative_attitude_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.

Wahrscheinlichkeitsverteilung: Multinomial

Verknüpfungsfunktion: Logit (kumulativ)[a]

a. Ziel: naturalness

b. Dieser Koeffizient wurde auf den Wert null gesetzt, da er redundant ist.

Dependent Variable Comprehensibility

Modellübersicht

Ziel		understandability
Messniveau		Ordinal
Wahrscheinlichkeitsverteilung		Multinomial
Verknüpfungsfunktion		Logit (kumulativ)
Informationskriterium	Akaike (korrig	28352,817
	Bayes	28357,769

Informationskriterien beruhen auf der -2 Log-Likelihood (28350,813) und dienen zum Modellvergleich. Modelle mit kleineren Werten für Informationskriterien weisen eine bessere Anpassung auf.

Zusammenfassung der Fallverarbeitung

	N	Prozent
Eingeschlossen	1104	100,00%
Ausgeschlossen	0	0,00%
Gesamtergebnis	1104	100,00%

Klassifikation

Gesamtprozent korrekt = 75,3%[a]

Beobachtet		Vorhergesagt: gar nicht verständlich	eher nicht verständlich	teils/teils	eher verständlich	sehr gut verständlich
gar nicht verständlich	Anzahl	0	0	1	1	0
	% in 'Beobachtet'	0,00%	0,00%	50,00%	50,00%	0,00%
eher nicht verständlich	Anzahl	0	0	4	5	0
	% in 'Beobachtet'	0,00%	0,00%	44,40%	55,60%	0,00%
teils/teils	Anzahl	0	0	13	39	9
	% in 'Beobachtet'	0,00%	0,00%	21,30%	63,90%	14,80%
eher verständlich	Anzahl	0	0	6	115	146
	% in 'Beobachtet'	0,00%	0,00%	2,20%	43,10%	54,70%
sehr gut verständlich	Anzahl	0	0	2	60	703
	% in 'Beobachtet'	0,00%	0,00%	0,30%	7,80%	91,90%

a. Ziel: understandability

Kovarianzparameter-Übersicht

Kovarianzparameter		
Residualeffekt		0
Zufällige Effek		1
Design-Matrix-Spalte	Feste Effekte	132
	Zufällige Effek	1[a]
Gemeinsame Subjekte		46

Gemeinsame Subjekte beruhen auf den Subjektspezifikationen für den Residualeffekt und die zufälligen Effekte und dienen dazu, die Daten aufzuteilen, um eine bessere

a. Dies ist die Anzahl an Spalten pro gemeinsamem Subjekt.

Block für zufällige Effekte 1

Block für zufällige Ef	Konstanter Term
Konstanter Term	2,87

Kovarianzstruktur: Skalierte Identität

Subjektspezifikation: ID

Zufälliger Effekt

Zufälliger Effekt Kovarianz	Schätzer	Standard Fehler	Z	Sig.	95% Konfidenzintervall: Unterer Wert	Oberer Wert
Varianz	2,87	1	2,87	0,004	1,45	5,682

Kovarianzstruktur: Skalierte Identität

Subjektspezifikation: ID

Feste Effekte[a]

Quelle	F	df1	df2	Sig.
Korrigiertes Modell	5136,701	45	1049	0,000
complexity	2,25	1	1049	0,134
domain	0,677	1	1049	0,411
sentence_type	98,569	1	1049	0,000
question_type	0,902	2	1049	0,406
complexity * sentence_type	0,117	1	1049	0,733
domain * sentence_type	5,032	1	1049	0,025
question_type * sentence_type	3,466	2	1049	0,032
age_groups	5,211	3	1049	0,001
gender	14,731	1	1049	0,000
experience_linguistics_kat	8,847	2	1049	0,000
experience_fas_kat	3,248	2	1049	0,039
experience_enc_kat	0,752	2	1049	0,472
bfi_openness_kat	2,336	1	1049	0,127
bfi_conscientiousness_kat	0,02	1	1049	0,888
bfi_extraversion_kat	3,527	1	1049	0,061
bfi_agreeableness_kat	1,473	1	1049	0,225
bfi_neuroticism_kat	2,797	2	1049	0,061
ta_competence_kat	0,622	1	1049	0,431
ta_enthusiasm_kat	4,135	2	1049	0,016

...

(continued)

ta_positive_attitude_kat	9,214	1	1049	0,002
ta_negative_attitude_kat	7,721	1	1049	0,006
age_groups * sentence_type	2,277	3	1049	0,078
gender * sentence_type	0,33	1	1049	0,566
experience_linguistics_kat * sentence_type	327,406	2	1049	0,000
experience_fas_kat * sentence_type	1,217	2	1049	0,296
experience_enc_kat * sentence_type	0,024	2	1049	0,976
bfi_openness_kat * sentence_type	2,092	1	1049	0,148
bfi_conscientiousness_kat * sentence_type	3,556	1	1049	0,060
bfi_extraversion_kat * sentence_type	2,189	1	1049	0,139
bfi_agreeableness_kat * sentence_type	28,13	1	1049	0,000
bfi_neuroticism_kat * sentence_type	1,96	2	1049	0,141
ta_competence_kat * sentence_type	0,001	1	1049	0,979
ta_enthusiasm_kat * sentence_type	2,173	2	1049	0,114
ta_positive_attitude_kat * sentence_type	0,148	1	1049	0,700
ta_negative_attitude_kat * sentence_type	0,509	1	1049	0,476

Wahrscheinlichkeitsverteilung: Multinomial

Verknüpfungsfunktion: Logit (kumulativ)[a]

a. Ziel: understandability

Feste Koeffizienten[a]

Modellterm		Koeffizient	Standard Fehler	t	Sig.	Unterer Wert	Oberer Wert	Exp(Coefficient)	Unterer Wert	Oberer Wert
Schwellenwert für understandabi	1	-13,039	2,3609	-5,523	0	-17,671	-8,406	2,17E-06	2,12E-08	0
	2	-11,267	2,3614	-4,771	0	-15,901	-6,634	0,00001278	1,24E-07	0,001
	3	-8,907	2,2098	-4,031	0	-13,243	-4,571	0	1,77E-06	0,01
	4	-6,075	2,1536	-2,821	0,005	-10,301	-1,85	0,002	0,00003359	0,157
complexity=1		0,393	0,2635	1,493	0,136	-0,124	0,91	1,482	0,884	2,485
complexity=2		0[b]	.	.	.	.	.	.	.	.
domain=1		0,143	0,273	0,522	0,601	-0,393	0,678	1,153	0,675	1,971
domain=2		0[b]	.	.	.	.	.	.	.	.
sentence_type=1		-0,881	1,6499	-0,534	0,594	-4,118	2,357	0,414	0,016	10,556
sentence_type=2		0[b]	.	.	.	.	.	.	.	.
question_type=1		-0,44	0,2391	-1,842	0,066	-0,909	0,029	0,644	0,403	1,029
question_type=2		-0,011	0,2512	-0,045	0,964	-0,504	0,482	0,989	0,604	1,619
question_type=3		0[b]	.	.	.	.	.	.	.	.
[complexity=1]*[sentence_type=1]		-0,137	0,3997	-0,342	0,733	-0,921	0,648	0,872	0,398	1,911
[complexity=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[complexity=2]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[complexity=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[domain=1]*[sentence_type=1]		-0,67	0,2986	-2,243	0,025	-1,256	-0,084	0,512	0,285	0,92
[domain=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[domain=2]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[domain=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[question_type=1]*[sentence_type=1]		0,837	0,3397	2,465	0,014	0,171	1,504	2,311	1,186	4,5
[question_type=2]*[sentence_type=1]		0,51	0,3096	1,648	0,1	-0,097	1,118	1,666	0,907	3,058
[question_type=3]*[sentence_type=1]		0[b]	.	.	.	.	.	.	.	.
[question_type=1]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[question_type=2]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
[question_type=3]*[sentence_type=2]		0[b]	.	.	.	.	.	.	.	.
age_groups=1		-1,39	0,8055	-1,726	0,085	-2,971	0,19	0,249	0,051	1,21
age_groups=2		-0,957	0,8182	-1,17	0,242	-2,563	0,648	0,384	0,077	1,913
age_groups=3		1,22	0,9748	1,252	0,211	-0,692	3,133	3,388	0,5	22,945
age_groups=4		0[b]	.	.	.	.	.	.	.	.
gender=1		-2,017	0,499	-4,043	0	-2,996	-1,038	0,133	0,05	0,354
gender=2		0[b]	.	.	.	.	.	.	.	.
experience_linguistics_kat=1		-3,887	1,0271	-3,784	0	-5,902	-1,871	0,021	0,003	0,154
experience_linguistics_kat=2		-1,384	0,5679	-2,436	0,015	-2,498	-0,269	0,251	0,082	0,764
experience_linguistics_kat=3		0[b]	.	.	.	.	.	.	.	.
experience_fas_kat=1		1,884	0,9184	2,052	0,04	0,082	3,687	6,583	1,086	39,91
experience_fas_kat=2		-0,082	0,6062	-0,136	0,892	-1,272	1,107	0,921	0,28	3,026
experience_fas_kat=3		0[b]	.	.	.	.	.	.	.	.
experience_enc_kat=1		-1,103	0,8416	-1,311	0,19	-2,754	0,548	0,332	0,064	1,731
experience_enc_kat=2		-1,031	0,945	-1,092	0,275	-2,886	0,823	0,356	0,056	2,277
experience_enc_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_openness_kat=2		0,542	0,5332	1,017	0,31	-0,504	1,588	1,72	0,604	4,895
bfi_openness_kat=3		0[b]	.	.	.	.	.	.	.	.
bfi_conscientiousness_kat=2		-0,359	1,5342	-0,234	0,815	-3,369	2,652	0,699	0,034	14,179
bfi_conscientiousness_kat=3		0[b]	.	.	.	.	.	.	.	.

...

(continued)

bfi_extraversion_kat=2	-0,865	0,7319	-1,181	0,238	-2,301	0,571	0,421	0,1	1,771
bfi_extraversion_kat=3	0[b]	.	.	.	.	.	.	.	.
bfi_agreeableness_kat=2	-0,103	0,8012	-0,128	0,898	-1,675	1,469	0,902	0,187	4,346
bfi_agreeableness_kat=3	0[b]	.	.	.	.	.	.	.	.
bfi_neuroticism_kat=1	-2,549	1,4312	-1,781	0,075	-5,357	0,26	0,078	0,005	1,296
bfi_neuroticism_kat=2	-1,629	1,2867	-1,266	0,206	-4,154	0,896	0,196	0,016	2,45
bfi_neuroticism_kat=3	0[b]	.	.	.	.	.	.	.	.
ta_competence_kat=2	-0,679	0,8919	-0,762	0,446	-2,43	1,071	0,507	0,088	2,917
ta_competence_kat=3	0[b]	.	.	.	.	.	.	.	.
ta_enthusiasm_kat=1	-1,898	1,1603	-1,635	0,102	-4,174	0,379	0,15	0,015	1,461
ta_enthusiasm_kat=2	0,908	1,0816	0,84	0,401	-1,214	3,031	2,48	0,297	20,713
ta_enthusiasm_kat=3	0[b]	.	.	.	.	.	.	.	.
ta_positive_attitude_kat=2	-1,39	0,5272	-2,636	0,009	-2,424	-0,355	0,249	0,089	0,701
ta_positive_attitude_kat=3	0[b]	.	.	.	.	.	.	.	.
ta_negative_attitude_kat=2	-2,849	0,9395	-3,032	0,002	-4,692	-1,005	0,058	0,009	0,366
ta_negative_attitude_kat=3	0[b]	.	.	.	.	.	.	.	.
[age_groups=1]*[sentence_type=1]	0,205	0,7578	0,271	0,787	-1,282	1,692	1,228	0,278	5,431
[age_groups=2]*[sentence_type=1]	0,878	0,6665	1,318	0,188	-0,429	2,186	2,407	0,651	8,901
[age_groups=3]*[sentence_type=1]	0,908	0,9288	0,978	0,329	-0,915	2,731	2,479	0,401	15,342
[age_groups=4]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[age_groups=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[age_groups=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[age_groups=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[age_groups=4]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[gender=1]*[sentence_type=1]	0,26	0,4518	0,575	0,566	-0,627	1,146	1,296	0,534	3,146
[gender=2]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[gender=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[gender=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=1]*[sentence_type=1]	14,183	0,5703	24,87	0	13,064	15,301	1443394,603	471428,004	4419313,162
[experience_linguistics_kat=2]*[sentence_type=1]	0,904	0,5195	1,741	0,082	-0,115	1,924	2,47	0,891	6,846
[experience_linguistics_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_linguistics_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=1]*[sentence_type=1]	0,713	0,591	1,207	0,228	-0,446	1,873	2,041	0,64	6,508
[experience_fas_kat=2]*[sentence_type=1]	0,028	0,4197	0,066	0,947	-0,796	0,851	1,028	0,451	2,342
[experience_fas_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_fas_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_enc_kat=1]*[sentence_type=1]	-0,156	1,1715	-0,133	0,894	-2,454	2,143	0,856	0,086	8,526
[experience_enc_kat=2]*[sentence_type=1]	-0,106	1,2717	-0,084	0,933	-2,602	2,389	0,899	0,074	10,903
[experience_enc_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[experience_enc_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_enc_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[experience_enc_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_openness_kat=2]*[sentence_type=1]	0,475	0,3283	1,446	0,148	-0,169	1,119	1,608	0,844	3,062
[bfi_openness_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_openness_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_openness_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_conscientiousness_kat=2]*[sentence_type=1]	1,18	0,6255	1,886	0,06	-0,048	2,407	3,253	0,953	11,098
[bfi_conscientiousness_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_conscientiousness_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_conscientiousness_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_extraversion_kat=2]*[sentence_type=1]	-0,988	0,6678	-1,48	0,139	-2,299	0,322	0,372	0,1	1,38
[bfi_extraversion_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_extraversion_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_extraversion_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_agreeableness_kat=2]*[sentence_type=1]	2,011	0,3791	5,304	0	1,267	2,754	7,468	3,549	15,712
[bfi_agreeableness_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_agreeableness_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_agreeableness_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=1]*[sentence_type=1]	0,425	0,8375	0,507	0,612	-1,218	2,068	1,529	0,296	7,911
[bfi_neuroticism_kat=2]*[sentence_type=1]	1,074	0,7551	1,422	0,155	-0,408	2,556	2,927	0,665	12,88
[bfi_neuroticism_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[bfi_neuroticism_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_competence_kat=2]*[sentence_type=1]	-0,015	0,555	-0,026	0,979	-1,104	1,074	0,985	0,332	2,928
[ta_competence_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_competence_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_competence_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=1]*[sentence_type=1]	-0,426	0,5803	-0,734	0,463	-1,565	0,713	0,653	0,209	2,04
[ta_enthusiasm_kat=2]*[sentence_type=1]	-1,091	0,5397	-2,021	0,044	-2,15	-0,032	0,336	0,117	0,969
[ta_enthusiasm_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=1]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_enthusiasm_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.

• • •

(continued)

[ta_positive_attitude_kat=2]*[sentence_type=1]	-0,177	0,4586	-0,385	0,7	-1,076	0,723	0,838	0,341	2,061
[ta_positive_attitude_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_positive_attitude_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_positive_attitude_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_negative_attitude_kat=2]*[sentence_type=1]	0,31	0,435	0,713	0,476	-0,543	1,164	1,364	0,581	3,202
[ta_negative_attitude_kat=3]*[sentence_type=1]	0[b]	.	.	.	.	.	.	.	.
[ta_negative_attitude_kat=2]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.
[ta_negative_attitude_kat=3]*[sentence_type=2]	0[b]	.	.	.	.	.	.	.	.

Wahrscheinlichkeitsverteilung: Multinomial

Verknüpfungsfunktion: Logit (kumulativ)[a]

a. Ziel: understandability

b. Dieser Koeffizient wurde auf den Wert null gesetzt, da er redundant ist.

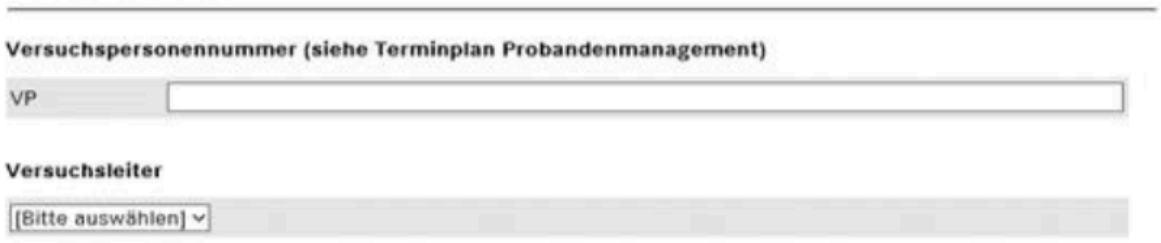
Appendix B

Materials of the Studies on Language Production

This appendix chapter contains the materials used for the data collection study for the purpose of investigating the linguistic behavior of drivers.

B.1 Pre-Survey

The purpose of this questionnaire was to measure individual user characteristics. Besides general demographic and linguistic-related questions, this survey includes the questionnaire by Rammstedt and Danner (2016) to measure Big Five Personality traits (s. “Persönlichkeits-bezogene Selbsteinschätzung"").



Demographische Daten

Alter

[ ] Jahre

Geschlecht

- ○ männlich
- ○ weiblich

Aktuelle Tätigkeit

- ○ nicht arbeitstätig
- ○ angestellt
- ○ selbstständig
- ○ in Ausbildung (Schule, Studium, etc.)
- ○ Hausfrau/Hausmann
- ○ Rentner/Pensionär
- ○ andere Tätigkeit

Persönlichkeitsbezogene Selbsteinschätzung

Inwieweit treffen die folgenden Aussagen auf Sie persönlich zu? Lesen Sie bitte jede Aussage aufmerksam durch und entscheiden Sie dann, wie sehr Sie der jeweiligen Aussage zustimmen. Bitte beantworten Sie jede Aussage spontan und wahrheitsgemäß.

Ich...	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
... bin gesprächig und unterhalte mich gern.	○	○	○	○	○
... neige dazu, andere zu kritisieren.	○	○	○	○	○
... erledige Aufgaben gründlich.	○	○	○	○	○
... bin deprimiert, niedergeschlagen.	○	○	○	○	○
... bin originell, entwickle neue Ideen.	○	○	○	○	○
... bin eher zurückhaltend, reserviert.	○	○	○	○	○
... bin hilfsbereit und selbstlos gegenüber anderen.	○	○	○	○	○
... bin manchmal unsorgfältig und schluderig.	○	○	○	○	○
... bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	○	○	○	○	○

...

Ich...	sehr un-zutreffend	eher un-zutreffend	teils/ teils	eher zutreffend	sehr zutreffend
... kann mich schroff und abweisend anderen gegenüber verhalten.	○	○	○	○	○
... mache Pläne und führe sie auch durch.	○	○	○	○	○
... werde leicht nervös und unsicher.	○	○	○	○	○
... stelle gerne Überlegungen an, spiele mit abstrakten Ideen.	○	○	○	○	○
... habe nur wenig künstlerisches Interesse.	○	○	○	○	○
... verhalte mich kooperativ, ziehe Zusammenarbeit dem Wettbewerb vor.	○	○	○	○	○
... bin leicht ablenkbar, bleibe nicht bei der Sache.	○	○	○	○	○
... kenne mich gut in Musik, Kunst oder Literatur aus.	○	○	○	○	○
... habe oft Krach mit anderen.	○	○	○	○	○

Wie bewerten Sie Ihre sprachliche Begabung?

Unter dem Begriff „Sprachbegabung" verstehen wir allgemein ein „Gefühl" für Sprache und grammatische Strukturen.

Die folgenden Fragen können Ihnen bei Ihrer Einschätzung helfen:

- Lesen und erzählen Sie gerne?
- Fällt es Ihnen leicht, neue Sprachen zu erlernen oder sich an früher gelernte Sprachen zu erinnern?
- Haben Sie Freude an Sprachspielen, Reimen, Mehrdeutigkeiten, Hintersinn von Wörtern und Redewendungen?
- Würden Sie Ihren Sprachschatz als umfangreich an Wörtern und Ausdrücken beschreiben? Fällt es Ihnen leicht, sich mitzuteilen ohne dass Ihnen die Wörter fehlen?
- Können Sie die grammatischen Regeln einer Sprache leicht anwenden?

Nicht gemeint sind konkrete Sprachkenntnisse, d. h. wie gut Sie beispielsweise Italienisch sprechen oder welche Note Sie im Deutschunterricht hatten.

	sehr niedrig				sehr hoch
Selbseinschätzung zur Sprachbegabung	○	○	○	○	○

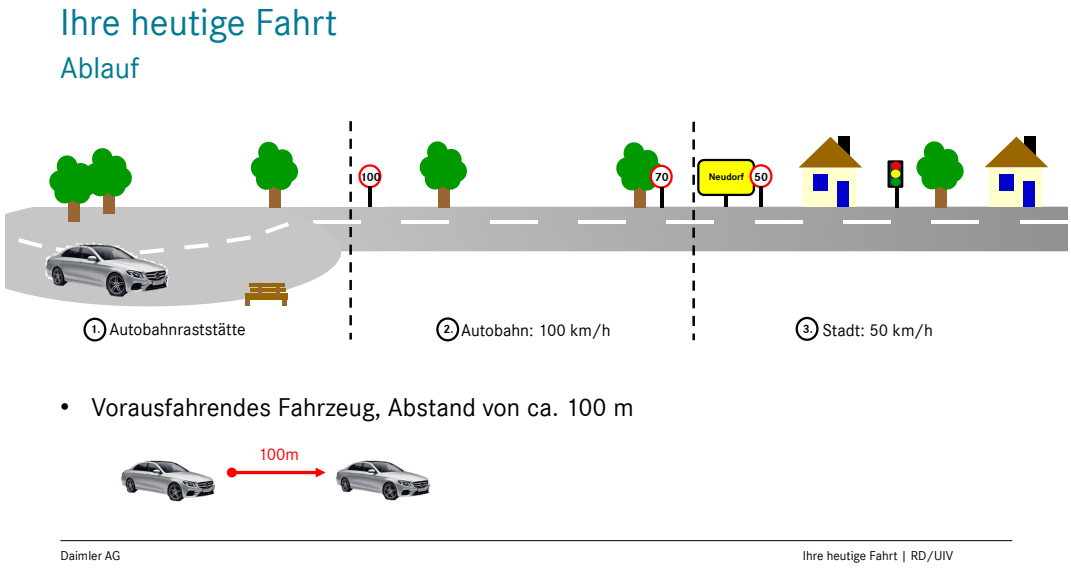
B.2 Small Talk Questions

Part		Question
parked position	1	Hallo! Das Wetter zur Zeit ist ja ziemlich gut. Wie gefällt Dir dieser Sommer und wie nutzt Du das Wetter?
	2	Wie gefällt Dir dieses Auto, wenn Du Dich etwas umsiehst? Was hättest Du gerne anders umgesetzt?
	3	Mein Lieblingsessen ist Lasagne. Was hälst du von Lasagne? Was ist dein Lieblingsessen?
	4	Wo warst Du das letzte mal im Urlaub und welche Erinnerung hast Du daran?
	5	Was würdest Du tun, wenn Du morgen im Lotto gewinnen würdest? Welche Wünsche würdest Du Dir gerne erfüllen?
	6	Was sind für Dich gute Freizeitaktivitäten für Deine Familie? Welche Erfahrungen hast Du da?
	7	Früher traditionell - heute eher viel und bunt: Wie wird in Deiner Familie der Weihnachtsbaum geschmückt? Welcher Schmuck und
	8	Welche Art von Urlaub machst Du lieber: Städtetrip oder Pauschalreise? Und warum?
	9	Welchen Beruf wolltest Du als Kind immer haben und aus welchem Grund?
	10	Was machst Du nach einem stressigen Arbeitstag? Wie kannst Du Dich entspannen?
	11	Wie viele Pflanzen hast Du zu Hause und welche?
	12	Welche Musik hörst Du am liebsten und was gefällt Dir daran besonders gut?
highway	13	Welchen Film hast Du zuletzt gesehen und wovon handelte er?
	14	Wie findest Du Kreuzfahrten? Welche Erfahrungen hast Du damit bisher gemacht?
	15	Was ist ein Projekt, dass du daheim schon lange einmal umsetzen möchtest und warum?
	16	Wie würdest Du den morgigen Tag gestalten, wenn Du einen Tag frei hättest? Welche Ideen hast Du dazu?
	17	Wo möchtest Du unbedingt einmal Deinen Urlaub verbringen und warum?
	18	Im Großraum Stuttgart kommt es oft Stau. Welche Erfahrungen hast Du? Wie oft stehst Du im Stau, und wo?
	19	Was machst Du üblicherweise im Feierabend? Welche Pläne hast Du heute?
	20	An heißen Tagen springe ich am liebsten in den See. Was machst Du, um Dich bei heißem Wetter abzukühlen und was für Tipps hast
	21	Hast du eine Lieblingsstadt? Welche Stadt ist das und was gefällt Dir an ihr so besonders gut?
	22	Welche Freizeitaktivitäten magst Du im Winter? Welche Erfahrungen hast Du da?
	23	Welche Fremdsprachen würdest Du gerne lernen und warum?
	24	Was ist dein Lieblings-Eisbecher und welche Eis-Sorte magst du gar nicht?
	25	Wohin war Deine letzte längere Autofahrt und zu welchem Zweck?
	26	Welche Früchte magst du gerne und warum isst du diese am liebsten?
	27	Was für ein Haustier hättest Du gern oder was für ein Haustier hast Du bei Dir zu Hause?
	28	Mit welcher Person würdest Du gerne mal für einen Tag die Rollen tauschen und warum?
	29	Welche Spiele spielst Du gerne? Welche Spiele magst Du gar nicht?
	30	Welche Erinnerungen hast Du an Deine Schulzeit? Welche Erfahrungen fallen Dir dazu ein?
	31	Wie planst Du üblicherweise einen Urlaub und wie bereitest Du ihn vor?
city	32	Wo fährst Du überall mit dem Auto hin? Warum fährst Du dort mit dem Auto hin?
	33	Wie verbringst Du gerne Dein Wochenende? Welche Pläne hast Du für das kommende Wochenende?
	34	Was ist dein Lieblingshobby und wie kam es zu diesem Hobby?
	35	Ende des Jahres wird bei meisten Silvester in besonderer Form gefeiert. Wie sieht dein Lieblingssilvesterabend aus? Mit wem feierst
	36	Wann ist Deine liebste Jahreszeit und warum?
	37	Geburtstage werden bei Kindern meist groß gefeiert - später nicht mehr so. Welche Traditionen hat Deine Familie? Wie feiert ihr welche Geburtstage?
	38	Welche Art von Sport betreibst Du aktuell? Welche Sportart gefällt Dir im Allgemeinen gut und Warum?
	39	Was war Dein schönster Urlaub bisher? Wo hast Du diesen verbracht und an welche Erlebnisse denkst Du besonders gern zurück?
	40	Spielst du ein Instrument, wenn ja welches? Welches Instrument würdest du gerne lernen und warum?
	41	Was für eine Sportart würdest Du wählen, wenn Du damit Dein Geld verdienen müsstest? Was glaubst Du würdest Du gut können?
	42	Wo wäre für Dich der schönste Ort zu leben, wenn Du frei wählen könntest? Was für Gedanken hast Du Dir dazu bisher gemacht?
	43	Welche Freizeitaktivitäten machst Du im Sommer am liebsten? Was machst Du gar nicht gern?
	44	Wenn du bis zum Ende deines Lebens nur noch ein bestimmtes Gericht essen dürftest, welches wäre das? Wie wird es zubereitet?
	45	Wo hast Du als Kind die meiste Zeit verbracht? Welche Erinnerungen hast Du da?
	46	Welche berühmte Person würdest Du gerne einmal treffen und aus welchem Grund?
	47	In welches Restaurant gehst Du am liebsten? Was magst Du gar nicht?
	48	Was war in der Schule Dein Lieblingsfach? Warum hat es Dir am besten gefallen und ist das immer noch so?
	49	Was hältst Du von Videospielen und welche Erfahrungen hast Du selbst damit?
	50	Was würdest du alles auf eine paradiesische Insel mitnehmen und warum?

B.3 Explanation of Study Content

This appendix section contains the material employed to explain the study content and procedure to the participants.

B.3.1 Study Procedure



Ihre heutige Fahrt

Sie haben heute die Möglichkeit einen Sprachassistenten kennenzulernen. Während der Fahrt wird der Sprachassistent mit Ihnen sprechen und Ihnen Fragen stellen. Uns interessieren hier Ihre Antworten. Sprechen Sie mit dem Sprachassistenten wie mit einem menschlichen Gegenüber. So kann das System lernen, wie Menschen sprechen und kann sich dadurch in Zukunft besser auf Sie einstellen. Dadurch leisten Sie einen wertvollen Beitrag zur Weiterentwicklung des Systems.

Ihre Fahrt wird Sie über verschiedene Streckenabschnitte führen:

- Ihr Fahrzeug ist zunächst auf einer Autobahnraststätte abgestellt. Hier beginnen wir im Stand.
- Auf der Autobahn fahren Sie bitte mit möglichst konstanter Geschwindigkeit von 100 km/h.
- In der Stadt fahren Sie bitte mit möglichst konstanter Geschwindigkeit von 50 km/h.
- In der Stadt wird Ihr vorausfahrendes Fahrzeug in eine Parkbucht parallel zur Straße einscheren. Stellen Sie Ihr Fahrzeug bitte ebenfalls auf diesen Parkplatz hinter Ihrem Führungsfahrzeug ab und parken.

Während Ihrer Fahrt folgen Sie einem vorausfahrenden Fahrzeug und halten bitte einen Abstand von ca. 100 m ein (entspricht zwei Leitpfosten).

Der Sprachassistent wird während Ihrer Fahrt jeweils eine Unterhaltung zu verschiedenen Themengebieten beginnen, beispielsweise zu Ihrem Musikgeschmack oder Erlebnissen während einer Reise. Sie können in die Unterhaltung einsteigen, indem Sie einfach auf die vom Sprachassistenten eingeleitete Frage antworten und Ihre Meinung oder Ihre Erfahrung äußern.

Wir bitten Sie zu berücksichtigen, dass es uns heute um Sie und Ihre Sprache geht. Um eine vertrautere Basis für Ihre Gespräche während der Fahrt zu erzeugen, wird Sie der Sprachassistent Duzen. Wir bitten Sie, sich von dieser vielleicht ungewohnten Anrede nicht irritieren zu lassen. Ihre Gespräche während der Fahrt werden außerdem als Frage-Antwort-Sequenzen ablaufen – der Sprachassistent stellt Ihnen Fragen, die Sie beantworten dürfen. Bitte lassen Sie sich nicht irritieren, wenn der Sprachassistent keinen direkten Bezug auf Ihre Antworten nimmt oder Ihnen nicht antwortet.

Es gibt keine richtigen oder falschen Antworten auf die Fragen des Sprachassistenten – formulieren Sie bitte frei spontane Antworten. Sie dürfen gerne so viel erzählen, wie Ihnen einfällt. Gerne dürfen Sie auch über die Fragen des Sprachassistenten hinaus berichten und bspw. in ein eigenes Thema überleiten. Die Fragen des Sprachassistenten sollen Ihnen hauptsächlich als Vorschläge für mögliche Gesprächsthemen dienen.

Für den Sprachassistenten ist hierbei nicht interessant *was* Sie ihm erzählen, sondern vielmehr *wie* Sie über verschiedene Themen sprechen. Ihre Antworten werden selbstverständlich sensibel behandelt und anonymisiert.

Bitte lassen Sie sich nicht hinsichtlich der Stimme oder Aussprachefehler des Sprachassistenten beeinflussen.

Während der gesamten Fahrt haben Sie Sprechkontakt zu Ihrer Versuchsleiterin/Ihrem Versuchsleiter. Sollten Sie sich zu irgendeinem Zeitpunkt nicht wohl fühlen, geben Sie bitte sofort Bescheid. Sie können die Fahrt dann zu jedem Zeitpunkt unterbrechen bzw. abbrechen.

Wir wünschen gute Fahrt!

B.3.2 Example Questions

By means of examples, the participants were introduced to the type of small talk questions they were asked during the simulation experiment. They were instructed to formulate spontaneous answers.

BEISPIELFRAGEN

1. **Ich mache gerne Urlaub am Strand und liege dann dort den ganzen Tag. Wie verbringst Du am liebsten einen Strandurlaub?**

Mögliche Antwort:

- Mir wird schnell langweilig, wenn ich nur am Strand liege.
- Letztes Mal am Strand habe ich Surfen gelernt... Surfen ist seitdem mein Lieblingshobby, ich habe mich schon für einen weiteren Kurs angemeldet und bin dabei mich vorzubereiten...

2. **Für viele Leute ist Gartenarbeit entspannend. Wie findest Du Gartenarbeit? Welche Erfahrungen hast Du damit?**

Mögliche Antwort:

- Ich selbst habe gar keinen Garten.
- Aber wir sind oft bei meinen Eltern zu Besuch, dann grillen wir oft gemeinsam. Bei gutem Wetter grillen wir sehr gerne und kommen alle zusammen... es ist schön, wenn die Familie zusammen kommt...

B.3.3 Avatar Selection

The participants were asked to choose an avatar as conversational partner during the experiment. The text samples below were synthesized and played to the participants to give them an impression about their options of either the female avatar Petra or the male avatar Yannick.

Presentation text: Petra

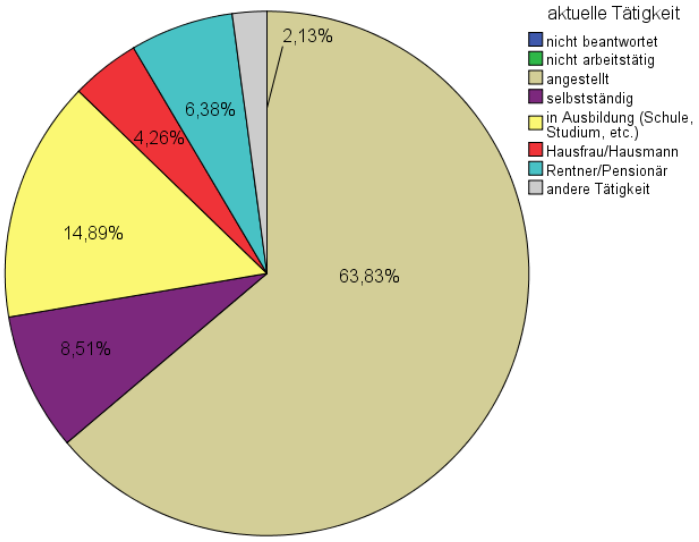
Hallo, ich bin Petra. Ich würde mich freuen, Dich heute auf Deiner Fahrt nach Neudorf begleiten zu dürfen. Dabei würde ich Dich gerne etwas näher kennen lernen. Was hältst Du davon?

Presentation text: Yannick

Hallo, mein Name ist Yannick. Wollen wir gleich gemeinsam nach Neudorf fahren? Auf der Fahrt würde ich Dich gerne etwas näher kennen lernen. Im Gespräch geht so eine Fahrt ja immer sehr viel schneller vorüber, nicht wahr?

B.4 Supplementary Analysis

This section contains supplementary results about the participants from the pre-survey. The majority of participants reported being in a salaried or apprenticeship position at the time of the study.



Appendix C

Materials for the Development of an Adaptive Strategy

This appendix chapter includes the required material for the development approach of a user- and situation-adaptive strategy concerning the syntactic form of voice output.

C.1 Principal Component Analysis

Deskriptive Statistiken

	Mittelwert	Std.-Abweichung	Analyse N
prop_count	19,40	20,480	1220
wordcount	69,76	78,567	1220
TTR	,812310221	,124195451	1220
stopwords_count	45,85	53,518	1220
max_syntactic_depth	6,49	2,398	1220
most_right_root_position	,415374707	,222902190	1220
max_verb_valence	2,36	,644	1220
max_dependency_length	15,53	8,487	1220
root_positions_per_word	,019614958	,013294587	1220
word_dependencies_per_word	,929060823	,039061046	1220
dependency_length_per_word	3,47613892	,601659210	1220
most_left_root_position	,123445171	,085531732	1220
max_root_dependencies	5,10	1,701	1220
root_dependencies_per_word	,277184029	,109507634	1220
max_word_dependency	5,60	1,518	1220
syntactic_depth_per_word	,304569946	,068398036	1220
complementizers_per_word	,021372362	,022753568	1220
modifiers_per_word	,272490047	,079726189	1220
pure_mainclauses_per_word	,248247420	,333075192	1220
relative_clauses_per_word	,011458154	,020817419	1220

Kommunalitäten

	Anfänglich	Extraktion
prop_count	1,000	,886
wordcount	1,000	,903
TTR	1,000	,776
stopwords_count	1,000	,897
max_syntactic_depth	1,000	,654
most_right_root_position	1,000	,619
max_verb_valence	1,000	,323
max_dependency_length	1,000	,599
root_positions_per_word	1,000	,831
word_dependencies_per_word	1,000	,567
dependency_length_per_word	1,000	,747
most_left_root_position	1,000	,383
max_root_dependencies	1,000	,785
root_dependencies_per_word	1,000	,743
max_word_dependency	1,000	,794
syntactic_depth_per_word	1,000	,683
complementizers_per_word	1,000	,738
modifiers_per_word	1,000	,485
pure_mainclauses_per_word	1,000	,755
relative_clauses_per_word	1,000	,780

Extraktionsmethode: Hauptkomponentenanalyse.

Korrelationsmatrix [a]

a. Determinante = 9,415E-10		prop_count	wordcount	TTR	stopwords_count	max_syntactic_depth	most_right_root_position	max_verb_valence	max_dependency_length	root_positions_per_word	word_dependencies_per_word	dependency_length_per_word	most_left_root_position	max_root_dependencies
Korrelation	prop_count	1,000	,983	-,763	,980	,461	,450	,431	,487	-,105	,053	,152	-,347	,471
	wordcount	,983	1,000	-,772	,998	,490	,446	,442	,509	-,128	,081	,175	-,352	,474
	TTR	-,763	-,772	1,000	-,770	-,497	-,486	-,442	-,538	,137	-,108	-,237	,395	-,541
	stopwords_count	,980	,998	-,770	1,000	,489	,439	,438	,506	-,129	,084	,178	-,351	,470
	max_syntactic_depth	,461	,490	-,497	,489	1,000	,153	,277	,492	-,452	,363	,371	-,408	,195
	most_right_root_position	,450	,446	-,486	,439	,153	1,000	,259	,212	,486	-,213	-,066	,038	,375
	max_verb_valence	,431	,442	-,442	,438	,277	,259	1,000	,339	-,157	,073	,237	-,293	,353
	max_dependency_length	,487	,509	-,538	,506	,492	,212	,339	1,000	-,353	,297	,639	-,389	,427
	root_positions_per_word	-,105	-,128	,137	-,129	-,452	,486	-,157	-,353	1,000	-,575	-,570	,517	-,094
	word_dependencies_per_word	,053	,081	-,108	,084	,363	-,213	,073	,297	-,575	1,000	,603	-,185	,059
	dependency_length_per_word	,152	,175	-,237	,178	,371	-,066	,237	,639	-,570	,603	1,000	-,301	,228
	most_left_root_position	-,347	-,352	,395	-,351	-,408	,038	-,293	-,389	,517	-,185	-,301	1,000	-,284
	max_root_dependencies	,471	,474	-,541	,470	,195	,375	,353	,427	-,094	,059	,228	-,284	1,000
	root_dependencies_per_word	-,089	-,125	,142	-,124	-,566	,159	-,110	-,334	,666	-,465	-,545	,253	,248
	max_word_dependency	,477	,491	-,562	,490	,313	,297	,387	,502	-,259	,180	,440	-,358	,833
	syntactic_depth_per_word	-,230	-,245	,352	-,243	-,115	-,062	-,260	-,509	,398	-,207	-,565	,210	-,447
	complementizers_per_word	-,031	-,012	-,050	-,003	,087	-,032	,031	,169	-,230	,202	,414	-,109	-,017
	modifiers_per_word	,151	,141	-,248	,143	,056	,123	,075	,203	-,135	,015	,245	-,135	,368
	pure_mainclauses_per_word	-,329	-,325	,429	-,326	-,280	-,300	-,277	-,340	,116	,041	-,292	,300	-,255
	relative_clauses_per_word	-,220	-,207	,248	-,200	-,021	-,279	-,115	,002	-,248	,192	,297	,038	-,256
Sig. (1-seitig)	prop_count		,000	,000	,000	,000	,000	,000	,000	,000	,032	,000	,000	,000
	wordcount	,000		,000	,000	,000	,000	,000	,000	,000	,002	,000	,000	,000
	TTR	,000	,000		,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	stopwords_count	,000	,000	,000		,000	,000	,000	,000	,000	,002	,000	,000	,000
	max_syntactic_depth	,000	,000	,000	,000		,000	,000	,000	,000	,000	,000	,000	,000
	most_right_root_position	,000	,000	,000	,000	,000		,000	,000	,000	,000	,010	,094	,000
	max_verb_valence	,000	,000	,000	,000	,000	,000		,000	,000	,005	,000	,000	,000
	max_dependency_length	,000	,000	,000	,000	,000	,000	,000		,000	,000	,000	,000	,000
	root_positions_per_word	,000	,000	,000	,000	,000	,000	,000	,000		,000	,000	,000	,001
	word_dependencies_per_word	,032	,002	,000	,002	,000	,000	,005	,000	,000		,000	,000	,020
	dependency_length_per_word	,000	,000	,000	,000	,000	,010	,000	,000	,000	,000		,000	,000
	most_left_root_position	,000	,000	,000	,000	,000	,094	,000	,000	,000	,000	,000		,000
	max_root_dependencies	,000	,000	,000	,000	,000	,000	,000	,000	,001	,020	,000	,000	
	root_dependencies_per_word	,001	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	max_word_dependency	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	syntactic_depth_per_word	,000	,000	,000	,000	,000	,015	,000	,000	,000	,000	,000	,000	,000
	complementizers_per_word	,141	,339	,040	,457	,001	,130	,136	,000	,000	,000	,000	,000	,277
	modifiers_per_word	,000	,000	,000	,000	,026	,000	,004	,000	,000	,301	,000	,000	,000
	pure_mainclauses_per_word	,000	,000	,000	,000	,000	,000	,000	,000	,000	,077	,000	,000	,000
	relative_clauses_per_word	,000	,000	,000	,000	,227	,000	,000	,477	,000	,000	,000	,091	,000

Korrelationsmatrix[a]

		root_dependencies_per_word	max_word_dependency	syntactic_depth_per_word	complementizers_per_word	modifiers_per_word	pure_mainclauses_per_word	relative_clauses_per_word
Korrelation	prop_count	-,089	,477	-,230	-,031	,151	-,329	-,220
	wordcount	-,125	,491	-,245	-,012	,141	-,325	-,207
	TTR	,142	-,562	,352	-,050	-,248	,429	,248
	stopwords_count	-,124	,490	-,243	-,003	,143	-,326	-,200
	max_syntactic_depth	-,566	,313	-,115	,087	,056	-,280	-,021
	most_right_root_position	,159	,297	-,062	-,032	,123	-,300	-,279
	max_verb_valence	-,110	,387	-,260	,031	,075	-,277	-,115
	max_dependency_length	-,334	,502	-,509	,169	,203	-,340	,002
	root_positions_per_word	,666	-,259	,398	-,230	-,135	,116	-,248
	word_dependencies_per_word	-,465	,180	-,207	,202	,015	,041	,192
	dependency_length_per_word	-,545	,440	-,565	,414	,245	-,292	,297
	most_left_root_position	,253	-,358	,210	-,109	-,135	,300	,038
	max_root_dependencies	,248	,833	-,447	-,017	,368	-,255	-,256
	root_dependencies_per_word	1,000	-,024	,258	-,296	,038	,217	-,309
	max_word_dependency	-,024	1,000	-,560	,121	,418	-,361	-,120
	syntactic_depth_per_word	,258	-,560	1,000	-,272	-,442	,356	-,118
	complementizers_per_word	-,296	,121	-,272	1,000	,065	-,489	,628
	modifiers_per_word	,038	,418	-,442	,065	1,000	-,218	-,051
	pure_mainclauses_per_word	,217	-,361	,356	-,489	-,218	1,000	-,411
	relative_clauses_per_word	-,309	-,120	-,118	,628	-,051	-,411	1,000
Sig. (1-seitig)	prop_count	,001	,000	,000	,141	,000	,000	,000
	wordcount	,000	,000	,000	,339	,000	,000	,000
	TTR	,000	,000	,000	,040	,000	,000	,000
	stopwords_count	,000	,000	,000	,457	,000	,000	,000
	max_syntactic_depth	,000	,000	,000	,001	,026	,000	,227
	most_right_root_position	,000	,000	,015	,130	,000	,000	,000
	max_verb_valence	,000	,000	,000	,136	,004	,000	,000
	max_dependency_length	,000	,000	,000	,000	,000	,000	,477
	root_positions_per_word	,000	,000	,000	,000	,000	,000	,000
	word_dependencies_per_word	,000	,000	,000	,000	,301	,077	,000
	dependency_length_per_word	,000	,000	,000	,000	,000	,000	,000
	most_left_root_position	,000	,000	,000	,000	,000	,000	,091
	max_root_dependencies	,000	,000	,000	,277	,000	,000	,000
	root_dependencies_per_word		,201	,000	,000	,095	,000	,000
	max_word_dependency	,201		,000	,000	,000	,000	,000
	syntactic_depth_per_word	,000	,000		,000	,000	,000	,000
	complementizers_per_word	,000	,000	,000		,012	,000	,000
	modifiers_per_word	,095	,000	,000	,012		,000	,038
	pure_mainclauses_per_word	,000	,000	,000	,000	,000		,000
	relative_clauses_per_word	,000	,000	,000	,000	,038	,000	

Erklärte Gesamtvarianz

	Anfängliche Eigenwerte			Summen von quadrierten Faktorladungen für Extraktion		Summen von quadrierten Faktorladungen für Extraktion	Rotierte Summe der quadrierten Ladungen[a]
Komponente	Gesamt	% der Varianz	Kumulierte %	Gesamt	% der Varianz	Kumulierte %	Gesamt
1	6,860	34,302	34,302	6,860	34,302	34,302	5,949
2	3,574	17,869	52,171	3,574	17,869	52,171	3,761
3	1,885	9,427	61,597	1,885	9,427	61,597	3,889
4	1,630	8,151	69,748	1,630	8,151	69,748	2,674
5	,944	4,722	74,470				
6	,803	4,013	78,483				
7	,750	3,751	82,234				
8	,676	3,381	85,616				
9	,573	2,865	88,480				
10	,466	2,328	90,808				
11	,407	2,033	92,841				
12	,344	1,722	94,563				
13	,292	1,462	96,025				
14	,220	1,102	97,127				
15	,188	,940	98,066				
16	,163	,815	98,881				
17	,126	,629	99,511				
18	,074	,368	99,879				
19	,022	,111	99,989				
20	,002	,011	100,000				

Extraktionsmethode: Hauptkomponentenanalyse.

a. Wenn Komponenten korreliert sind, können die Summen der quadrierten Ladungen nicht addiert werden, um eine Gesamtvarianz zu erhalten.

C.2 Voice Prompts as Comparison

The complexity factors of the voice prompts employed in the user study concerning the perception of language were used as a guideline for the classification of complexity factors of the identified user clusters. The table below provides the standardized feature values and the resulting complexity factors, which were calculated using the factor loadings according to Table 5.1 (p. 159) and the formula 5.1 (p. 162). As can be seen from this, higher factor values are generally assumed for the syntactically simpler main clause variant MCV than for the syntactically more complex relative clause variant RCV.

Table C.1: Complexity factors of voice prompts (N = 24).

Features / Factors	Voice Prompts	
	MCV	RCV
Sentence length	0.85	0.85
Stopwords (cound)	0.84	0.84
Propositions (count)	0.91	0.86
Type-token ratio	0.88	0.85
Syntactic cepth (max)	0.75	0.87
Root position (most right)	0.94	0.86
Verb valence (max)	0.43	0.15
Dependency length (max)	0.83	0.82
Root position (mean)	0.92	0.65
Word dependency (mean)	0.91	0.91
Root dependency (mean)	0.78	0.83
Dependency length (mean)	0.79	0.87
Root position (most left)	0.94	0.93
Root dependency (max)	0.80	0.80
Word dependency (max)	0.84	0.76
Modifiers (Dep. label)	0.78	0.69
Syntactic depth (mean)	0.85	0.72
Complementizers (Dep. label)	–	–
Relative clauses (Synt. structure)	0.98	0.98
Main clauses (Synt. structure)	0.97	0.97
Factor (1)	0.80	0.77
Factor (2)	0.91	0.53
Factor (3)	0.79	0.77
Factor (4)	0.09	0.09

C.3 Materials of the Real-Life User Study

C.3.1 Pre-Survey

The participants were guided through the preliminary and Big Five Personality questionnaires by Karrer *et al.* (2009) by means of a VBA-based interface. From the start screen, the participants could either begin the survey or access a help page including instructions. The last figure provides examples of the graphical interface asking the participant to enter age and gender.

This figure has been removed due to copyright limitations.

Figure C.1: The start screen providing the possibility to either “start” the questionnaire or go to an “instructions” page.

This figure has been removed due to copyright limitations.

Figure C.2: The instructions page providing hints how to navigate through the questionnaire.

These figures have been removed due to copyright limitations.

Figure C.3: Example screens of the questionnaire asking the participant to enter age (left) and gender (right). The participant was able to navigate through the quesitonnaire with the “back” and “next” buttons. Once all questions were answered, the “finish” button appeared as clickable to close the survey.

C.3.2 Intermediate and Post Survey

The purpose of this questionnaire was to measure individual driver distraction and experience with the presented dialog system. It includes the DALI questionnaire based on Hofmann (2015, s. part A) as well as the UEQ by Laugwitz *et al.* (2006, s. part B).

TEIL 1

A) Bitte bewerten Sie Ihre Erfahrung mit dem soeben erlebten Sprachassistenten. Bitte beziehen Sie Ihre Antwort dabei immer auf die Gesamtanforderungen, d. h. die Anforderungen, die Sie während Ihrer Fahrt in Verbindung mit dem Frage-Antwort-Dialog und der anschließenden Bewertung erlebt haben. Bitte beantworten Sie jede Aussage möglichst spontan.

	gering 1	2	3	4	5	6	hoch 7
1. Wie hoch waren die Anforderungen an die globale Aufmerksamkeit? Erklärung: Insgesamt alle mentalen - denken, entscheiden, … - visuellen und auditiven Faktoren, die insgesamt während des des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen.							
2. Wie hoch waren die visuellen Anforderungen? Erklärung: Visuelle Faktoren, die während des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen – alles, was mit dem Sehen zu tun hat.							
3. Wie hoch waren die auditiven Anforderungen? Erklärung: Auditive Faktoren, die während des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen – alles, was mit Gehörtem zu tun hat.							
4. Wie stark war das Stressniveau? Erklärung: Stressniveau während des Autofahrens mit Frage-Antwort-Dialogen wie Irritation, Müdigkeit, Unsicherheit, Entmutigung, etc.							
5. Wie hoch war die zeitliche Anforderung? Erklärung: Gefühlte Belastung und spezifische Beeinträchtigung durch die schnelle Abfolge der Aufgaben.							
6. Wie stark war der Interferenzfaktor? Erklärung: Die Fahrerbeanspruchung und ihre Auswirkung auf die Fahrleistung durch den gleichzeitig ablaufenden Frage-Antwort-Dialog während des Autofahrens.							

1

B) *Bitte bewerten Sie den soeben erlebten Sprachassistenten, indem Sie seine Eigenschaften mithilfe der Gegensatzpaare in den nachfolgenden Listen einschätzen. Entscheiden Sie dabei möglichst spontan, ohne lange über die Begriffe nachzudenken. Bitte kreuzen Sie immer eine Antwort an, auch wenn Sie bei der Einschätzung zu einem Begriffspaar unsicher sind. Es gibt keine „richtige" oder „falsche" Antwort. Ihre persönliche Meinung zählt!*

Bitte konzentrieren Sie sich bei der Bewertung auf die Interaktion mit dem System und die Form der System-Antworten. Versuchen Sie bitte, nicht die Qualität der Sprachausgabe zu bewerten.

	gering						hoch	
1.	1	2	3	4	5	6	7	
technisch								menschlich
unerfreulich								erfreulich
unverständlich								verständlich
kreativ								phantasielos
leicht zu lernen								schwer zu lernen
wertvoll								minderwertig
langweilig								spannend
uninteressant								interessant

	gering						hoch	
2.	1	2	3	4	5	6	7	
unberechenbar								voraussagbar
schnell								langsam
originell								konventionell
behindernd								unterstützend
gut								schlecht
kompliziert								einfach

	gering						hoch	
3.	1	2	3	4	5	6	7	
abstoßend								anziehend
herkömmlich								neuartig
unangenehm								angenehm
sicher								unsicher
aktivierend								einschläfernd
erwartungskonform								nicht erwartungskonform
ineffizient								effizient
übersichtlich								verwirrend

	gering						hoch	
4.	1	2	3	4	5	6	7	
unpragmatisch								pragmatisch
aufgeräumt								überladen
attraktiv								attraktiv
sympathisch								unsympathisch
konservativ								innovativ

TEIL 2

A) Bitte bewerten Sie Ihre Erfahrung mit dem soeben erlebten Sprachassistenten. Bitte beziehen Sie Ihre Antwort dabei immer auf die Gesamtanforderungen, d. h. die Anforderungen, die Sie während Ihrer Fahrt in Verbindung mit dem Frage-Antwort-Dialog und der anschließenden Bewertung erlebt haben. Bitte beantworten Sie jede Aussage möglichst spontan.

	gering 1	2	3	4	5	6	hoch 7
1. Wie hoch waren die Anforderungen an die globale Aufmerksamkeit? Erklärung: Insgesamt alle mentalen - denken, entscheiden, … - visuellen und auditiven Faktoren, die insgesamt während des des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen.							
2. Wie hoch waren die visuellen Anforderungen? Erklärung: Visuelle Faktoren, die während des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen – alles, was mit dem Sehen zu tun hat.							
3. Wie hoch waren die auditiven Anforderungen? Erklärung: Auditive Faktoren, die während des Autofahrens mit dem Frage-Antwort-Dialog erforderlich sind, um die Gesamtleistung zu erzielen – alles, was mit Gehörtem zu tun hat.							
4. Wie stark war das Stressniveau? Erklärung: Stressniveau während des Autofahrens mit Frage-Antwort-Dialogen wie Irritation, Müdigkeit, Unsicherheit, Entmutigung, etc.							
5. Wie hoch war die zeitliche Anforderung? Erklärung: Gefühlte Belastung und spezifische Beeinträchtigung durch die schnelle Abfolge der Aufgaben.							
6. Wie stark war der Interferenzfaktor? Erklärung: Die Fahrerbeanspruchung und ihre Auswirkung auf die Fahrleistung durch den gleichzeitig ablaufenden Frage-Antwort-Dialog während des Autofahrens.							

B) *Bitte bewerten Sie den soeben erlebten Sprachassistenten, indem Sie seine Eigenschaften mithilfe der Gegensatzpaare in den nachfolgenden Listen einschätzen. Entscheiden Sie dabei möglichst spontan, ohne lange über die Begriffe nachzudenken. Bitte kreuzen Sie immer eine Antwort an, auch wenn Sie bei der Einschätzung zu einem Begriffspaar unsicher sind. Es gibt keine „richtige" oder „falsche" Antwort. Ihre persönliche Meinung zählt!*

Bitte konzentrieren Sie sich bei der Bewertung auf die Interaktion mit dem System und die Form der System-Antworten. Versuchen Sie bitte, ***nicht*** *die Qualität der Sprachausgabe zu bewerten.*

	gering						hoch	
1.	1	2	3	4	5	6	7	
technisch								menschlich
unerfreulich								erfreulich
unverständlich								verständlich
kreativ								phantasielos
leicht zu lernen								schwer zu lernen
wertvoll								minderwertig
langweilig								spannend
uninteressant								interessant

	gering						hoch	
2.	1	2	3	4	5	6	7	
unberechenbar								voraussagbar
schnell								langsam
originell								konventionell
behindernd								unterstützend
gut								schlecht
kompliziert								einfach

	gering						hoch	
3.	1	2	3	4	5	6	7	
abstoßend								anziehend
herkömmlich								neuartig
unangenehm								angenehm
sicher								unsicher
aktivierend								einschläfernd
erwartungskonform								nicht erwartungskonform
ineffizient								effizient
übersichtlich								verwirrend

	gering						hoch	
4.	1	2	3	4	5	6	7	
unpragmatisch								pragmatisch
aufgeräumt								überladen
attraktiv								attraktiv
sympathisch								unsympathisch
konservativ								innovativ

C.3.3 Syntactic Paraphrases

This appendix section provides an overview of the syntactic paraphrases for the two domains DAS and COP. They were generated in the form of four levels with increasing syntactic complexity and employed in this real-life study as explanatory voice prompts as respective QAS answers.

C.3.4 Example Questions

BEISPIELFRAGEN

1.	**Was ist?**	Was ist der Nothalt-Assistent?
2.	**Wie funktioniert?**	Wie funktioniert der Nothalt-Assistent?
3.	**Wie ausschalten?**	Wie kann ich den Nothalt-Assistenten ausschalten?
4.	**Welche Arten?**	Welche Arten von Entertainment-Funktionen gibt es hier?
5.	**Ist vorhanden?**	Ist hier eine Entertainment-Funktion vorhanden?

Table C.2: Syntactic paraphrases generated for the domain DAS.

DAS function	Complexity level 1	Complexity level 2	Complexity level 3	Complexity level 4
Aufmerksamkeits-Assistent (*Attention Assist*)	Der Aufmerksamkeits-Assistent ist ein Sicherheitssystem. Er misst bei langen oder monotonen Fahrten deine Aufmerksamkeit. Der Aufmerksamkeits-Assistent warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit. Gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.	Der Aufmerksamkeits-Assistent ist ein Sicherheitssystem, das bei langen oder monotonen Fahrten deine Aufmerksamkeit misst. Der Aufmerksamkeits-Assistent warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit und gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.	Der Aufmerksamkeits-Assistent, der bei langen oder monotonen Fahrten deine Aufmerksamkeit misst, ist ein Sicherheitssystem. Er warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit und gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.	Der Aufmerksamkeits-Assistent, der als Sicherheitssystem bei langen oder monotonen Fahrten deine Aufmerksamkeit misst, warnt dich durch einen Signalton bei nachlassender Aufmerksamkeit und gleichzeitig erscheint im Kombiinstrument die Empfehlung für eine baldige Pause.
Brems-Assistent (*Brake Assist*)	Der eingeschaltete Brems-Assistent warnt dich mit einem Piepton vor einer drohenden Kollision. Er zeigt dir zusätzlich eine Warnung im Kombiinstrument. Der Brems-Assistent kann bei einer ausbleibenden Reaktion zu deinem Schutz selbständig bremsen.	Der eingeschaltete Brems-Assistent warnt dich mit einem Piepton vor einer drohenden Kollision und zeigt dir zusätzlich eine Warnung im Kombiinstrument. Der Brems-Assistent kann bei einer ausbleibenden Reaktion zu deinem Schutz selbständig bremsen.	Der eingeschaltete Brems-Assistent, der dich mit einem Piepton vor einer drohenden Kollision warnt, zeigt dir zusätzlich eine Warnung im Kombiinstrument. Er kann bei einer ausbleibenden Reaktion zu deinem Schutz selbständig bremsen.	Der eingeschaltete Brems-Assistent, der dich mit einem Piepton vor einer drohenden Kollision warnt, zeigt dir zusätzlich eine Warnung im Kombiinstrument und kann bei einer ausbleibenden Reaktion zu deinem Schutz selbständig bremsen.
Totwinkel-Assistent (*Blind Spot Assist*)	Der eingeschaltete Totwinkel-Assistent überwacht den toten Winkel deines Fahrzeugs. Er warnt dich bei einem gefundenen Hindernis durch eine rote Warnleuchte im Außenspiegel. Der Totwinkel-Assistent warnt dich vor einem Spurwechsel zusätzlich durch ein akustisches Signal.	Der eingeschaltete Totwinkel-Assistent überwacht den toten Winkel deines Fahrzeugs und warnt dich bei einem gefundenen Hindernis durch eine rote Warnleuchte im Außenspiegel. Der Totwinkel-Assistent warnt dich vor einem Spurwechsel zusätzlich durch ein akustisches Signal.	Der eingeschaltete Totwinkel-Assistent, der den toten Winkel deines Fahrzeugs überwacht, warnt dich bei einem gefundenen Hindernis durch eine rote Warnleuchte im Außenspiegel. Er warnt dich vor einem Spurwechsel zusätzlich durch ein akustisches Signal.	Der eingeschaltete Totwinkel-Assistent, der den toten Winkel deines Fahrzeugs überwacht, warnt dich bei einem gefundenen Hindernis durch eine rote Warnleuchte im Außenspiegel und vor einem Spurwechsel zusätzlich durch ein akustisches Signal.

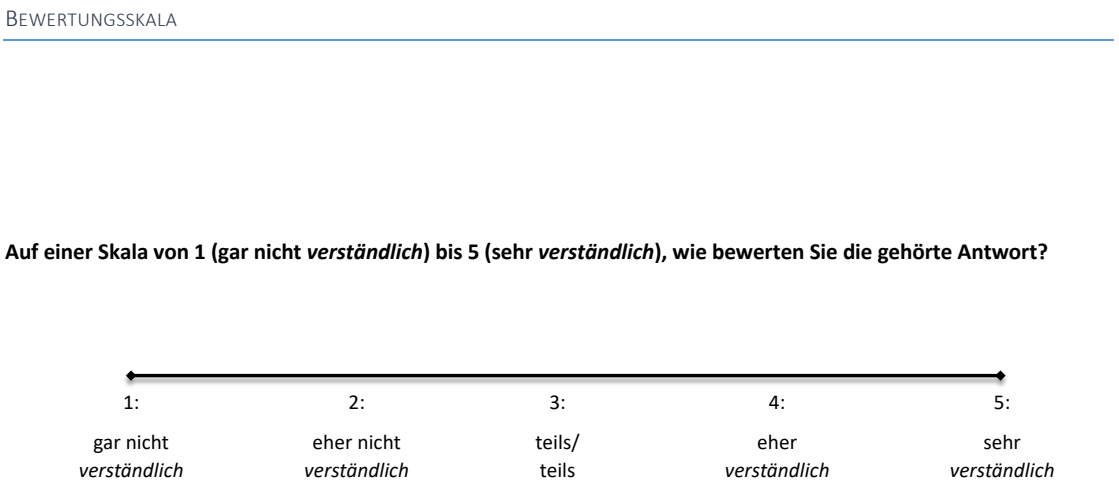
Table C.3: Syntactic paraphrases generated for the domain COP.

COP function	Complexity level 1	Complexity level 2	Complexity level 3	Complexity level 4
Beduftung (*Perfume atomizer*)	Das Beduftungssystem verteilt auf Wunsch einen angenehmen Duft im Fahrzeuginnenraum. Eine Auswahl an verschiedenen Parfümen steht hierbei zur Verfügung. Das Beduftungssystem kann so unterschiedliche Geruchseindrücke erzeugen. Es kann dadurch zum Wohlbefinden der Fahrzeuginsassen beitragen.	Das Beduftungssystem verteilt auf Wunsch einen angenehmen Duft im Fahrzeuginnenraum, wofür eine Auswahl an verschiedenen Parfümen zur Verfügung steht. Das Beduftungssystem kann so unterschiedliche Geruchseindrücke erzeugen und dadurch zum Wohlbefinden der Fahrzeuginsassen beitragen.	Das Beduftungssystem, wofür eine Auswahl an verschiedenen Parfümen zur Verfügung steht, verteilt auf Wunsch einen angenehmen Duft im Fahrzeuginnenraum. Es kann so unterschiedliche Geruchseindrücke erzeugen und dadurch zum Wohlbefinden der Fahrzeuginsassen beitragen.	Das Beduftungssystem, das auf Wunsch einen angenehmen Duft aus einer Auswahl an verschiedenen verfügbaren Parfümen im Fahrzeuginnenraum verteilt, kann so unterschiedliche Geruchseindrücke erzeugen und dadurch zum Wohlbefinden der Fahrzeuginsassen beitragen.
Massage-Programme (*Massage functions*)	Dieses Fahrzeug hat mehrere Massageprogramme. Sie können dein Wohlbefinden auf vielfältige Weise steigern. Zum Beispiel gibt dir das Massageprogramm Klassik mit lokalen Druckpunkten und Wellenbewegungen eine entspannende Massage. Ein weiteres Beispiel ist das Hot-Relaxing-Massageprogramm. Es massiert dir mit Wärme und punktuellem Druck deinen Rücken entlang.	Dieses Fahrzeug hat mehrere Massageprogramme und können dein Wohlbefinden auf vielfältige Weise steigern. Zum Beispiel gibt dir das Massageprogramm Klassik mit lokalen Druckpunkten und Wellenbewegungen eine entspannende Massage. Ein weiteres Beispiel ist das Hot-Relaxing-Massageprogramm, das dir mit Wärme und punktuellem Druck deinen Rücken entlang massiert.	Dieses Fahrzeug hat mehrere Massageprogramme, die dein Wohlbefinden auf vielfältige Weise steigern können. Zum Beispiel das Massageprogramm Klassik, das dir mit lokalen Druckpunkten und Wellenbewegungen eine entspannende Massage gibt. Ein weiteres Beispiel ist das Hot-Relaxing-Massageprogramm, das dir mit Wärme und punktuellem Druck deinen Rücken entlang massiert.	Dieses Fahrzeug hat mehrere Massageprogramme, die dein Wohlbefinden auf vielfältige Weise steigern können. Zum Beispiel das Massageprogramm Klassik, das dir mit lokalen Druckpunkten und Wellenbewegungen eine entspannende Massage gibt, oder das Hot-Relaxing-Massageprogramm, das dir mit Wärme und punktuellem Druck deinen Rücken entlang massiert.
Sprach-Assistent (*Linguatronic*)	Du kannst den MBUX Sprach-Assistenten nicht deaktivieren. Er hört aber nur in zwei Situationen auf deine Sprachbefehle. Zum Beispiel sagst du Hey Mercedes oder drückst die Taste mit dem stilisierten Kopf am Lenkrad.	Du kannst den MBUX Sprach-Assistenten nicht deaktivieren, aber er hört nur in zwei Situationen auf deine Sprachbefehle. Zum Beispiel sagst du Hey Mercedes oder drückst die Taste mit dem stilisierten Kopf am Lenkrad.	Der MBUX Sprach-Assistent, den du nicht deaktivierenk annst, hört aber nur in zwei Situationen auf deine Sprachbefehle. Zum Beispiel sagst du Hey Mercedes oder drückst die Taste mit dem stilisierten Kopf am Lenkrad.	Der MBUX Sprach-Assistent, den du nicht deaktivieren kannst, hört aber nur in zwei Situationen auf deine Sprachbefehle, indem du Hey Mercedes sagst oder die Taste mit dem stilisierten Kopf am Lenkrad drückst.

Table C.4: Standard (*i.e.*, non-adaptive) voice prompts generated for the domain DAS & COP.

DAS function	**MCV**
Spurhalte-Assistent (*Lane Keeping Assist*)	Der eingeschaltete Spurhalte-Assistent warnt dich beim Berühren einer Fahrbahnmarkierung. Er schützt dich so vor einem ungewollten Verlassen deiner Fahrspur. Der Spurhalte-Assistent führt dich beim Berühren einer durchgezogenen Markierung zurück auf deine Spur.
Verkehrszeichen-Assistent (*Traffic Assist*)	Der eingeschaltete Verkehrszeichen-Assistent warnt dich mit einem Piepton vor zu schnellem Fahren. Er zeigt dir zusätzlich ein Geschwindigkeitsschild im Kombiinstrument an. Der Verkehrszeichen-Assistent kann dich auch beim Fahren entgegen der vorgeschriebenen Richtung warnen.
Innovationen (*Innovation features*)	Dein Fahrzeug verfügt über eine Vielzahl an technischen Innovationen. Sie entsprechen dem Stand modernster Technik. Zum Beispiel versteht dich der MBUX SprachAssistent dank Künstlicher Intelligenz und kann viele Systeme deines Fahrzeugs bedienen. Ein weiteres Beispiel ist der Lenk-Assistent. Er kann mit leichten Lenkeingriffen helfen dein Fahrzeug in der Spur zu halten.
Diebstahl-Warnanlage (*Anti-Theft Alarm System*)	Ja, dieses Fahrzeug hat eine Einbruch-Diebstahl-Warnanlage. Die Einbruch-Diebstahl-Warnanlage wird mit der Verriegelung deines Fahrzeugs von außen automatisch aktiviert.
COP function	**MCV**
Lordosenstütze (*Lumbar pad*)	Die Lordosenstütze ist eine ergonomische Rückenstütze im Sitz. Sie ist der natürlichen Krümmung der Wirbelsäule nachempfunden. Die Lordosenstütze fördert das gesunde Sitzen. Sie trägt so bei langen Autofahrten zum Komfort der Fahrzeuginsassen bei.
4D-Tiefenmassage (*4D Sound Deep Massage*)	Du kannst die 4D-Tiefenmassage im Energizing Comfort-Programm Freude nutzen. Das 4D-Soundsystem erzeugt dabei mit Körperschallwandlern Vibrationsimpulse in der Sitzlehne. Die 4D-Tiefenmassage begünstigt so die positive Stimmung der Fahrzeuginsassen. Sie kann ebenso deren mentale Regeneration fördern.

C.3.5 Evaluation Scale



C.3.6 Distribution of User Clusters

Table C.5: Distribution of user clusters (UCs) and their characteristics.

		UC 1	UC 2	UC 3	$\sum$
	Subjects	1	2	3	6
	Age (SD)	55 (0)	35.5 (5.5)	47.67 (16.78)	44.83 (15.54)
	Gender	m: 0, f: 1	m: 1, f: 1	m: 2, f: 1	m: 3, f: 3
	Mileage (SD)	9,000 (0)	15,000 (5,000)	13,333 (4,714)	13,166 (5,307)
Big Five trait	**Agreeableness**	3.40 (0)	3.85 (0.35)	3.73 (0.09)	3.72 (0.26)
	Conscientiousness	4.78 (0)	3.06 (0.28)	3.67 (0.48)	3.65 (0.69)
	Extraversion	3.50 (0)	3.19 (0.81)	3.04 (0.62)	3.17 (0.66)
	Neuroticism	1.25 (0)	3.13 (0.63)	2.83 (0.59)	2.67 (0.85)
	Openness	3.70 (0)	2.40 (0.80)	2.80 (0.29)	2.82 (0.67)

Note: m – male; f – female;

C.3.7 Distribution of Voice Prompts and User Ratings

Table C.6 provides details concerning the distribution of user ratings and syntactic variants per user cluster for both systems *ADAPT* and *STAND*. In the case of UC 3, the additional differentiation between highway and city is made given the different adaptation behavior according to the strategy depicted in Figure 5.3 (p. 164).

Table C.6: Distribution of user ratings and syntactic variant per user cluster and driving situation for *ADAPT* and *STAND*.

	UC	H/C	Complexity level	User rating 1	2	3	4	5	Σ	*Comp*	M_{COMP}
ADAPT	1	H & C	1	0	0	0	0	0	0	–	4.50
			2	1	0	0	0	3	4	4.00	
			3	0	0	0	0	2	2	5.00	
			4	0	0	0	0	0	0	–	
	2	H & C	1	0	0	0	0	2	2	5.00	4.74
			2	0	0	0	1	4	5	4.80	
			3	0	0	0	1	2	3	4.67	
			4	0	0	0	1	1	2	4.50	
	3	H	1	0	0	0	0	0	0	–	3,67
			2	0	1	0	0	0	1	2.00	
			3	0	0	0	3	1	4	4.25	
			4	0	0	0	1	3	4	4.75	
		C	1	0	0	1	3	1	5	4.00	4,25
			2	0	0	1	0	3	4	4.50	
			3	0	0	0	0	0	0	–	
			4	0	0	0	0	0	0	–	
STAND	1	H & C	1	0	0	2	1	3	6	4.17	4.17
	2	H & C	1	0	0	2	3	7	12	4.42	4.42
	3	H & C	1	0	0	2	5	11	18	4.50	4.50

Note: UC – User Cluster; H – Highway; C – City.

Bibliography

Abdulla, W. H., Kasabov, N. K., and Zealand, D. (2001). Improving speech recognition performance through gender separation. *Changes*, **9**, 10.

Allen, J. (1995). *Natural Language Understanding*. The Benjamins / Cummings Publishing Company, Redwood City, CA.

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2001). Toward conversational human-computer interaction. *AI Magazine*, **22**(4), 27–27.

Aly, A. and Tapus, A. (2016). Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction. *Autonomous Robots*, **40**(2), 193–209.

Angkititrakul, P., Kwak, D., Choi, S., Kim, J., PhucPhan, A., Sathyanarayana, A., and Hansen, J. H. L. (2007). Getting start with UTDrive: Driver-behavior modeling and assessment of distraction for in-vehicle speech systems. In *8th Annual Conference of the International Speech Communication Association*, pages 1334–1337. ISCA.

Bach, K. M., Jæger, M. G., Skov, M. B., and Thomassen, N. G. (2009). Interacting with in-vehicle systems: Understanding, measuring, and evaluating attention. In *Proceedings of the 23rd Annual Conference on People and Computers: Celebrating People and Technology*, pages 453–462. ACM.

Bader, M. (2015). Leseverstehen und Sprachverarbeitung. (German). [Reading comprehension and language processing.]. In *Lesen: Ein interdisziplinäres Handbuch*, pages 141–168. Walter de Gruyter, Berlin/Boston.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley Framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. ACL.

Barón, A. and Green, P. (2006). Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review. Technical report, University of Michigan, Transportation Research Institute Ann Arbor, MI.

Bartsch, R. (1973). *Gibt es einen sinnvollen Begriff von linguistischer Komplexität? (German). [Is there a meaningful notion of linguistic complexity?]*. De Gruyter, Berlin/New York.

Basil, M. D. (2012). Multiple resource theory. In *Encyclopedia of the Sciences of Learning*, pages 2384–2385. Springer Science & Business Media.

Bavelas, J. B., Black, A., Lemery, C. R., and Mullett, J. (1986). "I show how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, **50**(2), 322–329.

Becic, E., Dell, G. S., Bock, K., Garnsey, S. M., Kubose, T., and Kramer, A. F. (2010). Driving impairs talking. *Psychonomic Bulletin & Review*, **17**(1), 15–21.

Bell, A. (1984). Language style as audience design. *Language in Society*, **13**(2), 145–204.

Bell, L. (2003). *Linguistic Adaptations in Spoken Human-Computer Dialogues – Empirical Studies of User Behavior*. Ph.D. thesis, KTH Computer Science and Communication, University of Stockholm.

Berg, M. (2013). Natürlichsprachlichkeit in Dialogsystemen. (German) [Natural language in dialog systems.]. *Informatik-Spektrum*, **36**(4), 371–381.

Bergmann, K., Branigan, H. P., and Kopp, S. (2015). Exploring the alignment space – lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*, **2**, 7.

Bernsen, N. O., Dybkjær, H., and Dybkjær, L. (2012). *Designing interactive speech systems: From first ideas to user testing*. Springer Science & Business Media.

Birkner, K. (2008). *Relativ(satz)konstruktionen im gesprochenen Deutsch. Syntaktische, prosodische, semantische und pragmatische Aspekte. (German) [Relative clause constructions in spoken German. Syntactic, prosodic, semantic and pragmatic approaches.]*. De Gruyter, Berlin/New York.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, **18**(3), 355–387.

Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., and Kubose, T. T. (2007). Car talk,

car listen. In *Automaticity and Control in Language Processing*, pages 21–42. Psychology Press.

Boersma, P. and Weenink, D. (2020). Praat: Doing phonetics by computer [Computer program, version 6.1.09]. `http://www.praat.org/`.

Bradac, J. J., Mulac, A., and House, A. (1988). Lexical diversity and magnitude of convergent versus divergent style shifting: Perceptual and evaluative consequences. *Language & Communication*, **8**(3-4), 213–228.

Branham, S. M. and Mukkath Roy, A. R. (2019). Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International Conference on Computers and Accessibility*, pages 446–458. ACM.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, **75**(2), 13–25.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Nass, C. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, volume 31, pages 186–191. Lawrence Erlbaum Associates.

Branigan, H. P., Pickering, M. J., Pearson, J., and McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, **42**(9), 2355–2368.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**(6), 1482.

Brouwer, S., Mitterer, H., and Huettig, F. (2010). Shadowing reduced speech and alignment. *The Journal of the Acoustical Society of America*, **128**(1), 32–37.

Brüggemeier, B., Breiter, M., Kurz, M., and Schiwy, J. (2020). User Experience of Alexa when controlling music: Comparison of face and construct validity of four questionnaires. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–9. ACM.

Burnett, G. E. and Ditsikas, D. (2006). Personality as a criterion for selecting usability testing participants. In *Proceedings of the International Conference on Information and Communications Technologies*, pages 599–604. IEEE.

Cavedon, L., Weng, F., Mishra, R., Bratt, H., Raghunathan, B., Cheng, H., Schmidt, H.,

Mirkovic, D., Bei, B., Pon-Barry, H., *et al.* (2005). Developing a conversational in-car dialog system. In *International Conference on Intelligent Transport Systems*. IEEE.

Chand, V., Baynes, K., Bonnici, L., and Farias, S. T. (2012). Analysis of idea density (AID): A manual. *University of California at Davis*.

Chen, F., Jonsson, I.-M., Villing, J., and Larsson, S. (2010). Application of speech technology in vehicles. In *Speech Technology*, pages 195–219. Springer.

Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Sigkdd Explorations Newsletter*, **19**(2), 25–35.

Chen, J., Bangalore, S., Rambow, O., and Walker, M. A. (2002). Towards automatic generation of natural language generation systems. In *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1, pages 1–7. ACL.

Chen, T., Huang, C., Chang, E., and Wang, J. (2001). Automatic accent identification using gaussian mixture models. In *Workshop on Automatic Speech Recognition and Understanding, 2001*, pages 343–346. IEEE.

Chomsky, N. (1965 [2014]). *Aspects of the theory of syntax*, volume 11. MIT Press.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, **22**(1), 1–39.

Cohen, J. (1992). A power primer. *Psychological bulletin*, **112**(1), 155.

Cohen, M. H., Cohen, M. H., Giangola, J. P., and Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.

Costa, P. T. and McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological assessment*, **4**(1), 5.

Costa, P. T. and McCrae, R. R. (1999). A five-factor theory of personality. *The Five-Factor Model of Personality: Theoretical Perspectives*, **2**, 51–87.

Covington, M. A. (2009). Idea density – A potentially informative characteristic of retrieved documents. In *Southeastcon*, pages 201–203. IEEE.

Dahlbäck, N. and Jönsson, A. (2007). Dialogue systems when the dialogue is just a secondary

task – Some preliminaries to the development of in-car dialogue systems. *Communication - Action - Meaning: A Festschrift to Jens Allwood*, **425**.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. ACL.

De Saussure, F. (1959 [2011]). *Course in general linguistics*. Columbia University Press.

De Waard, D. (1996). *The measurement of drivers' mental workload*. Ph.D. thesis, University of Groningen, Traffic Research Center Netherlands.

Deemter, K. v., Theune, M., and Krahmer, E. (2005). Real versus template-based natural language generation: A false opposition? *Computational linguistics*, **31**(1), 15–24.

Demberg, V. and Sayeed, A. (2011). Linguistic cognitive load: Implications for automotive UIs. *Workshop on Cognitive load and In-Vehicle Human-Machine Interaction*, pages 176–183.

Demberg, V., Winterboer, A., and Moore, J. D. (2011). A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, **37**(3), 489–539.

Demberg, V., Sayeed, A., Mahr, A., and Müller, C. (2013). Measuring linguistically-induced cognitive load during driving using the ConTRe task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 176–183. ACM.

Demberg, V., Hoffmann, J., Howcroft, D. M., Klakow, D., and Torralba, A. (2016). Search challenges in natural language generation with complex optimization objectives. *KI – Künstliche Intelligenz*, **30**(1), 63–69.

Dijkstra, P. and Barelds, D. P. H. (2008). Do people know what they want: A similar or complementary partner? *Evolutionary Psychology*, **6**(4), 595–602.

Durupınar, F., Pelechano, N., Allbeck, J., and Badler, N. (2011). The impact of the OCEAN personality model on the perception of crowds. *Computer Graphics and Applications*, **3**, 22–31.

Edlund, J., Gustafson, J., Heldner, M., and Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication*, **50**(8-9), 630–645.

Engström, J., Markkula, G., Victor, T., and Merat, N. (2017). Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human Factors*, **59**(5), 734–764.

Fast, L. A. and Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, **94**(2), 334.

Fellbaum, K. (2012). Sprachdialogsysteme (German) [Spoken dialog systems]. In *Sprachverarbeitung und Sprachübertragung*, pages 369–389. Springer.

Field, A. (2009). *Discovering statistics through SPSS*. Sage Publications, London, 3rd edition.

Fischer, G. (2001). User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction*, **11**(1), 65–86.

Florencio, E., Amores, G., Pérez, G., and Manchón, P. (2008). Aggregation in the in-home domain. *Procesamiento del lenguaje natural*, (40), 17–26.

Fowler, C. A., Brown, J. M., Sabadini, L., and Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, **49**(3), 396–413.

Franke, T., Attig, C., and Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human-Computer Interaction*, **35**(6), 456–467.

Fraser, N. M. and Gilbert, G. N. (1991). Simulating speech systems. *Computer Speech & Language*, **5**(1), 81–99.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, **27**(2), 181–218.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, **68**(1), 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, **2000**, 95–126.

Gill, A. J. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24. Taylor & Francis Group.

Goldberg, L. R. (1990). An alternative “description of personality”: The big-five factor structure. *Journal of personality and social psychology*, **59**(6), 1216.

Gordon-Salant, S. and Cole, S. S. (2016). Effects of age and working memory capacity on speech recognition performance in noise among listeners with normal hearing. *Ear and Hearing*, **37**(5), 593–602.

Green, P. (2001). Variations in task performance between younger and older drivers: UMTRI research on telematics. In *Association for the Advancement of Automotive Medicine Conference on Aging and Driving*. AAAM.

Grice, H. P. (1975). Logic and conversation. In *Speech Acts*, pages 41–58. Brill.

Grothkopp, D., Krautter, W., Grothkopp, B., Steffens, F., and Geutner, F. (2001). Using a driving simulator to perform a Wizard-of-Oz experiment on speech-controlled driver information systems. In *Proceedings of the 1st Human-Centered Transportation Simulation Conference*. Citeseer.

Hamerich, S. (2010). *Sprachbedienung im Automobil: Teilautomatisierte Entwicklung benutzerfreundlicher Dialogsysteme (German) [Voice control in the automobile: Partially automated development of user-friendly dialog systems]*. Springer.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, Advances in Psychology, pages 139–183. North-Holland Press.

Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität (german) [AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In *Mensch & Computer: Interaktion in Bewegung*, pages 187–196. B. G. Teubner.

Heck, R. H., Thomas, S., and Tabata, L. (2013). *Multilevel modeling of categorical outcomes using IBM SPSS*. Routledge Academic.

Heinroth, T. and Minker, W. (2012). *Introducing Spoken Dialogue Systems into Intelligent Environments*. Springer Science & Business Media.

Heisterkamp, P. (2001). Linguatronic: Product-level speech system for Mercedes-Benz car. In *Proceedings of the 1st International Conference on Human Language Technology Research*. ACL.

Hjalmarsson, A. (2005a). Adaptive spoken dialogue systems. In *Graduate School of Language Technology, Speech Technology, University of Göteborg*, pages 1–12.

Hjalmarsson, A. (2005b). Towards user modelling in conversational dialogue systems: A qualitative study of the dynamics of dialogue parameters. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 869–872. ISCA.

Hofmann, H. (2015). *Intuitive speech interface technology for information exchange tasks*. Ph.D. thesis, University of Ulm.

Hone, K. S. and Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, **6**(3-4), 287–303.

Honnibal, M. and Montani, I. (2017). SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., and Brown, J. (2006). Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis & Prevention*, **38**(1), 185–191.

Hu, Z., Tree, J. F., and Walker, M. A. (2018). Modeling linguistic and personality adaptation for natural language generation. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 20–31. ACL.

Huang, C., Chen, T., and Chang, E. (2004). Accent issues in large vocabulary continuous speech recognition. *International Journal of Speech Technology*, **7**(2), 141–153.

Hudson, R. (1995). *Measuring syntactic difficulty*. Unpublished manuscript. University College, London. Accessible via `https://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf` (06/06/2022).

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, **53**(2), 251–267.

Jain, J. J. and Busso, C. (2011). Analysis of driver behaviors during common tasks using frontal video camera and CAN-bus information. In *Proceedings of the International Conference on Multimedia and Expo*. IEEE.

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). Big five inventory. *Journal of Personality and Social Psychology*.

Jokinen, K. (2003). Natural interaction in spoken dialogue systems. In *Proceedings of the Workshop on Ontologies and Multilinguality in User Interfaces at HCI International*, pages 730–734. ACL.

Jokinen, K., Kerminen, A., Lagus, T., Kuusisto, J., Wilcock, G., Turunen, M., Hakulinen, J., and Jauhiainen, K. (2002). Adaptive dialogue systems – interaction with interact. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*, pages 64–73. ACL.

Jokinen, K., Kanto, K., and Rissanen, J. (2004). Adaptive user modelling in Athosmail. In *ERCIM Workshop on User Interfaces for All: User-Centered Interaction Paradigms for Universal Access in the Information Society*, pages 149–158. ACM.

Jones, E., Gallois, C., Callan, V., and Barker, M. (1999). Strategies of accommodation: Development of a coding system for conversational interaction. *Journal of Language and Social Psychology*, **18**(2), 123–151.

Jurafsky, D. and Martin, J. H. (2014). *Speech and Language Processing. An Introduction into Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, US, 3rd edition.

Just, M. A., Keller, T. A., and Cynkar, J. (2008). A decrease in brain activation associated with driving when listening to someone speak. *Brain Research*, **1205**, 70–80.

Karrer, K., Glaser, C., Clemens, C., and Bruder, C. (2009). Technikaffinität erfassen – der Fragebogen TA-EG. (German) [Measuring technical affinity – the questionnaire TA-EG]. *Der Mensch im Mittelpunkt technischer Systeme*, **8**, 196–201.

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K., and Mitzner, T. L. (2001). Language decline across the life span: Findings from the nun study. *Psychology and Aging*, **16**(2), 227.

Kern, D. and Schmidt, A. (2009). Design space for driver-based automotive user interfaces. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 3–10. ACM.

Kintsch, W. and Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, **5**(3), 257–274.

Kipp, M. (2003). *Gesture generation by imitation: From human behavior to computer character animation*. Ph.D. thesis, University of Saarland.

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, Melbourne, Australia. ACL.

Klatt, D. H. (1987). Review of text-to-speech conversion for english. *The Journal of the Acoustical Society of America*, **82**(3), 737–793.

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. In *Proceedings of the 13th Annual Symposium on User Interface Software and Technology*, pages 1–10. ACM.

Koch, P. (1995). Subordination, intégration syntaxique et “oralité”. (French) [Subordination, syntactic integration and “orality”]. *Études romanes*, **34**, 13–42.

Koch, P. and Oesterreicher, W. (2011). *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch. (German) [Spoken Language in Romania: French, Italian, Spanish]*, volume 31. Walter de Gruyter.

Kubose, T. T., Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., and Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, **20**(1), 43–63.

Lamel, L., Minker, W., and Paroubek, P. (2000). Towards best practice in the development and evaluation of speech recognition components of a spoken language dialog system. *Natural Language Engineering*, **6**(3-4), 305–322.

Landesberger, J., Ehrlich, U., and Minker, W. (2020). Do the urgent things first! – Detecting urgency in spoken utterances based on acoustic features. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. ACM.

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.

Large, D. R., Burnett, G., and Clark, L. (2019). Lessons from Oz: Design guidelines for automotive conversational user interfaces. In *Adjunct Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 335–340. ACM.

Laugwitz, B., Schrepp, M., and Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten (German) [construction of a questionnaire to measure the user experience of software products]. In *Mensch & Computer: Mensch und Computer im Strukturwandel*, pages 125–134. Oldenbourg Verlag.

Le Bigot, L., Terrier, P., Amiel, V., Poulain, G., Jamet, E., and Rouet, J.-F. (2007). Effect of modality on collaboration with a dialogue system. *International Journal of Human-Computer Studies*, **65**(12), 983–991.

Lee, K. M., Peng, W., Jin, S.-A., and Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, **56**(4), 754–772.

Lemon, O. (2008). Adaptive natural language generation in dialogue using reinforcement learning. In *Proceedings of the 12th SEMdial Workshop on the Semantics and Pragmatics of Dialogues*, pages 149–156. SEMDIAL.

Levelt, W. J. M. and Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, **14**(1), 78–106.

Levitan, S. I., Mishra, T., and Bangalore, S. (2016). Automatic identification of gender from speech. In *Proceeding of Speech Prosody*, pages 84–88. ISCA.

Li, P., Merat, N., Zheng, Z., Markkula, G., Li, Y., and Wang, Y. (2018). Does cognitive distraction improve or degrade lane keeping performance? Analysis of time-to-line crossing safety margins. *Transportation Research (Part F): Traffic Psychology and Behaviour*, **57**, 48–58.

Linell, P. (1998). *Approaching dialogue: Talk, interaction and contexts in dialogical perspectives*, volume 3. John Benjamins Publishing.

Litman, D. J. and Pan, S. (1999). Empirically evaluating an adaptable spoken dialogue system. In *Proceedings of the 7th International Conference on User Modeling*, pages 55–64.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, **9**(2), 159–191.

Liu, H., Xu, C., and Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, **21**, 171–193.

López-Cózar, R., Callejas, Z., Griol, D., and Quesada, J. F. (2014). Review of spoken dialogue systems. *Loquens: Revista Española de Ciencias del Habla*, **1**(2), e012.

Mairesse, F. and Walker, M. A. (2007). PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503. ACL.

Mairesse, F. and Walker, M. A. (2010). Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, **20**(3), 227–278.

Mairesse, F. and Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, **37**(3), 455–488.

Mairesse, F. and Young, S. (2014). Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, **40**(4), 763–799.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, **30**, 457–500.

Maybury, M. T. (2004). *New Directions in Question Answering*. AAAI Press.

McCrae, R. R. and Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, **52**(5), 509–516.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

McTear, M. F. (1993). User modelling for adaptive computer systems: A survey of recent developments. *Artificial Intelligence Review*, **7**(3), 157–184.

McTear, M. F. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys*, **34**(1), 90–169.

McTear, M. F. (2004). *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer Science & Business Media.

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. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 237–246. ACM.

Metze, F., Black, A., and Polzehl, T. (2011). A review of personality in voice-based man machine interaction. In *International Conference on Human-Computer Interaction*, pages 358–367. ACM.

Möller, S. (2004). *Quality of Telephone-Based Spoken Dialogue Systems*. Springer Science & Business Media.

Möller, S. (2017). *Quality Engineering: Qualität kommunikationstechnischer Systeme (German) [Quality Engineering: Quality of communication systems]*. Springer.

Moon, Y. and Nass, C. (1996). How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication Research*, **23**(6), 651–674.

Moore, J., Foster, M. E., Lemon, O., and White, M. (2004). Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*, pages 917–922. AAAI.

Moore, R. K. (2017). Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*, pages 281–291. Springer.

Muckler, F. A. and Seven, S. A. (1992). Selecting performance measures: "Objective" versus "subjective" measurement. *Human factors*, **34**(4), 441–455.

Muthig, J. and Schäflein-Armbruster, R. (2008). Funktionsdesign® – Methodische Entwicklung von Standards. *Standardisierungsmethoden für die Technische Dokumentation*, **16**, 41–74.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, **43**(2), 223–239.

Navarretta, C. (2016). Mirroring facial expressions and emotions in dyadic conversations. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 469–474. ACL.

Nielsen, F. and Minker, W. (2017). Assistive and adaptive dialog management. In *Companion Technology*, pages 167–186. Springer.

Nunes, L. and Recarte, M. A. (2002). Cognitive demands of hands-free-phone conversation while driving. *Transportation Research (Part F): Traffic Psychology and Behaviour*, **5**(2), 133–144.

Oh, A. and Rudnicky, A. (2000). Stochastic language generation for spoken dialogue systems. In *Workshop on Conversational Systems*, pages 27–31. ACL.

Oviatt, S., Darves, C., and Coulston, R. (2004). Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *Transactions on Computer-Human Interaction*, **11**(3), 300–328.

Papangelis, A., Karkaletsis, V., and Huang, H. (2013). Towards adaptive dialogue systems for assistive living environments. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 29–32. ACM.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, **119**(4), 2382–2393.

Paunonen, S. V. and Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, **81**(3), 524.

Pauzié, A. (2008). A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems*, **2**(4), 315–322.

Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., and Nass, C. I. (2006). Adaptive language behavior in HCI: How expectations and beliefs about a system affect users' word choice. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 1177–1180. ACM.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, **27**(2), 169–190.

Pieraccini, R. (2012). *The Voice in the Machine: Building Computers that Understand Speech*. MIT Press.

Pinter, Y., Reichart, R., and Szpektor, I. (2016). Syntactic parsing of web queries with question intent. In *Proceedings on Human Language Technologies*, pages 670–680. ACL.

Potamianos, A., Narayanan, S., and Lee, S. (1997). Automatic speech recognition for children. In *Proceedings of th 5th European Conference on Speech Communication and Technology*, pages 2371–2374. ISCA.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*, volume 14. Pearson Prentice Hall, US.

Rakauskas, M. E., Gugerty, L. J., and Ward, N. J. (2004). Effects of naturalistic cell phone conversations on driving performance. *Journal of Safety Research*, **35**(4), 453–464.

Rambow, O., Bangalore, S., and Walker, M. A. (2001). Natural language generation in dialog systems. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–4. ACL.

Rammstedt, B. and Danner, D. (2016). Die Facettenstruktur des Big Five Inventory (BFI). (German) [The multifaceted structure of the Big Five Inventory (BFI)]. *Diagnostica*, **63**(1), 70–84.

Ranney, T. A., Harbluk, J. L., and Noy, Y. I. (2005). Effects of voice technology on test track driving performance: Implications for driver distraction. *Human Factors*, **47**(2), 439–454.

Ratnaparkhi, A. (2000). Trainable methods for surface natural language generation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 194–201. ACL.

Raux, A., Langner, B., Bohus, D., Black, A. W., and Eskenazi, M. (2005). Let's go public! Taking a spoken dialog system to the real world. In *Proceedings of the International Speech Communication Association*. ISCA.

Reid, G. B., Shingledecker, C. A., and Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. In *Proceedings of the Human Factors Society Annual Meeting*, volume 25, pages 522–526. Sage Publications: Los Angeles, CA.

Reiter, E. (1994). Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 163–170. ACL.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, **3**(1), 57–87.

Reitter, D., Keller, F., and Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, **35**(4), 587–637.

Rieser, V., Lemon, O., and Liu, X. (2010). Optimising information presentation for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1009–1018. ACL.

Rizzolatti, G. and Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Current Opinion in Neurobiology*, **18**(2), 179–184.

Roth-Berghofer, T. R. and Cassens, J. (2005). Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In *International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development*, volume 3620, pages 451–464. Springer.

Russell, M. and D'Arcy, S. (2007). Challenges for computer recognition of children's speech. In *Workshop on Speech and Language Technology in Education*, pages 108–111. ISCA.

Ryckman, R. M. (2012). *Theories of Personality*. Cengage Learning, 10th edition.

SAE (2018). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard J3016201806, Rev 20180615. SAE International.

Scherer, K. R. and Scherer, U. (1981). Speech behavior and personality. *Speech Evaluation in Psychiatry*, **1**, 460.

Schmitt, A. and Minker, W. (2012). *Towards adaptive spoken dialog systems*. Springer Science & Business Media.

Skantze, G. (2007). *Error handling in spoken dialogue systems – Managing uncertainty, grounding and miscommunication*. Ph.D. thesis, KTH Computer Science and Communication, University of Stockholm.

Sørmo, F. and Cassens, J. (2004). Explanation goals in case-based reasoning. In *Proceedings of the ECCBR Workshops*, pages 165–174.

Sørmo, F., Cassens, J., and Aamodt, A. (2005). Explanation in case-based reasoning – Perspectives and goals. *Artificial Intelligence Review*, **24**(2), 109–143.

Spieker, P. (1991). *Natürlichsprachliche Erklärungen in technischen Expertensystemen (German) [Natural language explanations in technical expert systems]*. Ph.D. thesis, University of Kaiserslautern.

Stent, A., Prasad, R., and Walker, M. (2004). Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 79–86. ACL.

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. In *Pro-*

ceedings of the 27th Conference on User Modeling, Adaptation and Personalization, pages 191–195. ACM.

Stier, D., Heid, U., and Minker, W. (2020a). Adapting in-vehicle voice output: A user- and situation-adaptive approach. In *Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 12–15. ACM.

Stier, D., Heid, U., Kittel, P., Schmidt, M., and Minker, W. (2020b). The influence of syntax on the perception of in-vehicle prompts and driving performance. In *Conversational Dialogue Systems for the Next Decade*, pages 349–362. Springer.

Stier, D., Munro, K., Heid, U., and Minker, W. (2020c). Personality traits, speech and adaptive in-vehicle voice output. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, page 17–22. ACM.

Stier, D., Minker, W., and Heid, U. (2020d). Towards situation- and user-adaptive voice output: Classifying driver personality in context. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue – Short Papers*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.

Stier, D., Munro, K., Heid, U., and Minker, W. (2020e). Towards situation-adaptive in-vehicle voice output. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–7. ACM.

Stoyanchev, S. and Stent, A. (2009). Lexical and syntactic adaptation and their impact in deployed spoken dialog systems. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Short Papers*, pages 189–192. ACL.

Strauss, P.-M. and Minker, W. (2010). *Proactive Spoken Dialogue Interaction in Multi-Party Environments*. Springer.

Strayer, D. L. and Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, **12**(6), 462–466.

Strayer, D. L., Drews, F. A., and Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, **9**(1), 23.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., and Biondi, F. (2015a). Assessing cognitive distraction in the automobile. *Human Factors*, **57**(8), 1300–1324.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., and Hopman, R. J. (2015b). Measuring cognitive distraction in the automobile III: A comparison of ten 2015 in-vehicle information systems. *AAA Foundation for Traffic Safety*.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., and Hopman, R. J. (2016). Talking to your car can drive you to distraction. *Cognitive Research: Principles and Implications*, **1**(1), 1–16.

Suzuki, N. and Katagiri, Y. (2007). Prosodic alignment in human–computer interaction. *Connection Science*, **19**(2), 131–141.

Thomas, P., Czerwinski, M., McDuff, D., Craswell, N., and Mark, G. (2018). Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 42–51. ACM.

Turner, A. and Greene, E. (1977). *The Construction and Use of a Propositional Text Base*. Institute for the Study of Intellectual Behavior, University of Colorado Boulder.

Ultes, S., Kraus, M., Schmitt, A., and Minker, W. (2015). Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 374–383. ACL.

Van Baaren, R. B., Holland, R. W., Steenaert, B., and Van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, **39**(4), 393–398.

Varges, S. (2006). Overgeneration and ranking for spoken dialogue systems. In *Proceedings of the 4th International Natural Language Generation Conference*, pages 20–22. ACL.

Vergin, R., Farhat, A., and O'Shaughnessy, D. (1996). Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification. In *Proceeding of 4th International Conference on Spoken Language Processing*, volume 2, pages 1081–1084. IEEE.

Villing, J. (2009). In-vehicle dialogue management – Towards distinguishing between different types of workload. In *Proceedings of the 4th Workshop SiMPE on Speech in Mobile and Pervasive Environments*, pages 14–21. ACM.

Vogels, J., Demberg, V., and Kray, J. (2018). The index of cognitive activity as a measure of cognitive processing load in dual task settings. *Frontiers in Psychology*, **9**, 2276.

Vollrath, M. and Totzke, I. (2003). Möglichkeiten der Nutzung unterschiedlicher Ressourcen für die Fahrer-Fahrzeug-Interaktion (German) [Using multiple resources for the driver-vehicle-interaction]. *VDI-Berichte*, **1768**, 47–58.

Wachtel, S. (2003). Schreiben fürs Hören: Trainingstexte, Regeln und Methoden (German) [Writing for listening: training texts, rules and methods]. *Praktischer Journalismus*, **3**.

Walker, M. A., Whittaker, S. J., Stent, A., Maloor, P., Moore, J., Johnston, M., and Vasireddy, G. (2004). Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, **28**(5), 811–840.

Wang, Y., Reitter, D., and Yen, J. (2014). Linguistic adaptation in conversation threads: Analyzing alignment in online health communities. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 55–62. ACL.

Wärnestål, P. and Kronlid, F. (2014). Towards a user experience design framework for adaptive spoken dialogue in automotive contexts. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, pages 305–310. ACM.

Warren, T. and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, **85**(1), 79–112.

Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D., and Young, S. (2015a). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721. ACL.

Wen, T.-H., Gasic, M., Kim, D., Mrksic, N., Su, P.-H., Vandyke, D., and Young, S. (2015b). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284. ACL.

Weng, F., Angkititrakul, P., Shriberg, E. E., Heck, L., Peters, S., and Hansen, J. H. (2016). Conversational in-vehicle dialog systems: The past, present, and future. *Signal Processing Magazine*, **33**(6), 49–60.

White, M., Rajkumar, R., and Martin, S. (2007). Towards broad coverage surface realization with CCG. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation*, pages 267–276. ACL.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, **3**(2), 159–177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, **50**(3), 449–455.

Wickens, C. D. (2020). Processing resources and attention. In *Multiple-Task Performance*, pages 3–34. CRC Press.

Wierwille, W. W. (1993). Demands on driver resources associated with introducing advanced technology into the vehicle. *Transportation Research (Part C): Emerging Technologies*, **1**(2), 133–142.

Willemyns, M., Gallois, C., Callan, V. J., and Pittam, J. (1997). Accent accommodation in the job interview: Impact of interviewer accent and gender. *Journal of Language and Social Psychology*, **16**(1), 3–22.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, **3**(1), 1–191.

Xu, Y. and Reitter, D. (2015). An evaluation and comparison of linguistic alignment measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–67. ACL.

Xu, Y. and Reitter, D. (2016). Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 443–448. ACL.

Young, K., Regan, M., and Hammer, M. (2007). Driver distraction: A review of the literature. *Distracted Driving*, **206**, 379–405.

Young, K. L., Regan, M. A., and Lee, J. D. (2009). *Measuring the effects of driver distraction: Direct driving performance methods and measures*, pages 85–105. CRC Press.

Young, S., Gašić, M., Thomson, B., and Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. volume 101, pages 1160–1179. IEEE.

Yu, D. and Deng, L. (2016). *Automatic Speech Recognition*, volume 1. Springer.

List of Contributing Publications

Book Publications

1. D. Stier, U. Heid, P. Kittel, M. Schmidt, and W. Minker
   The influence of syntax on the perception of in-vehicle prompts and driving performance
   Conversational Dialogue Systems for the Next Decade, Springer Singapore, pp. 349–362, 2020
   Nominated for the Best Paper Award

2. M. Schmidt, D. Stier, S. Werner, and W. Minker
   Exploration and assessment of proactive use cases for an in-car voice assistant
   Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019, TUDpress Dresden, pp. 148–155, 2019

Conference Publications

1. D. Stier, U. Heid, and W. Minker
   Adapting in-vehicle voice output: A user- and situation-adaptive approach
   12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20), Virtually in Washington (D.C., USA), September 2020

2. D. Stier, K. Munro, U. Heid, and W. Minker
   Towards situation-adaptive in-vehicle voice output
   Proceedings of the 2nd ACM Conference on Conversational User Interfaces (ACM CUI '20), Virtually in Bilbao (Spain), July 2020

3. D. Stier, W. Minker, and U. Heid
   Towards Situation- and User-Adaptive Voice Output: Classifying Driver Personality in Context
   Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL '20 – Short Papers), Virtually at Brandeis (Waltham, New Jersey), July 2020

4. D. Stier, K. Munro, U. Heid, and W. Minker
   Personality traits, speech and adaptive in-vehicle voice output
   Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (ACM UMAP '20), Virtually in Genoa (Italy), July 2020

5. D. Stier and E. Sigloch
   Linguistic design of in-vehicle prompts in adaptive dialog systems: An analysis of potential factors involved in the perception of naturalness
   Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (ACM UMAP '19), Larnaca (Cyprus), June 2019

6. M. Schmidt, D. Helbig, O. Bhandare, D. Stier, W. Minker, and S. Werner
   Assessing objective indicators of users' cognitive load during proactive in-car dialogs
   Adjunct Publication of the 27th ACM Conference on User Modeling, Adaptation and Personalization (ACM UMAP '19 Adjunct), Larnaca (Cyprus), June 2019